FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

First order optimality conditions under state constraints

Igor Kornienko

November 2014

Faculty of Engineering, University of Porto, Portugal

Doctoral Program in Electrical and Computer Engineering

Abstract

The main focus of this thesis is on pure state and mixed state-control constraints for optimal control problems with a particular emphasis on first order conditions of optimality. It concerns both theory and applications. The applications are based on two problems for the control of infectious diseases involving a compartmental SEIR model. These are treated both analytically and computationally. Beyond their own value, they play a crucial role in this thesis as they allow us to illustrate some important theoretic concepts and results, and moreover, they are the drive behind our new and more theoretical results.

The first SEIR problem we consider has an L^2 cost and pure state constraints. Our second problem has an L^1 cost and mixed control-state constraints. The computational solutions of both problems are partially validated using the Maximum Principle. In this respect, normality and regularity conditions play an important role. We are not treating a particular disease in a specific population. Rather, we illustrate how different optimal control formulations can be used to propose new vaccination policies when different scenarios and cost functionals are considered.

Our analytical treatment of the solution to the L^2 state constrained problem motivates the study of exact penalization for problems with such constraints. We define a set of assumptions under which the proposed exact penalization scheme yields regularity of the multipliers.

Next, we apply necessary conditions for nonsmooth problems with mixed constraints to problems involving differential algebraic equations. Remarkably, we do not need to apply the computationally expensive implicit function theorem and we cover some problems with nonsmooth data.

Finally, we characterize control problems involving mixed state-control problems which can be described by differential inclusions. We establish conditions under which properties of the set-valued mapping defining the differential inclusion enable us to use well-known results in the literature for differential inclusion control problems.

Resumo

Esta tese concentra-se em problemas de controlo óptimo com restrições do estado puras e restrições mistas, em particular, nas condições necessárias de optimalidade da primeira ordem para estes problemas. Tratámos não só questões teóricas como aplicações. As aplicações de interesse nesta tese são baseadas em dois problemas do controlo de doenças infecciosas que involvem um modelo compartimental SEIR. Estes problemas são tratados analiticamente e computacionalmente. Tais problemas têm uma grande importância nesta tese, porque permitem-nos ilustrar alguns resultados e conceitos teóricos importantes. Além disso, eles dão o impulso para alguns dos nossos resultados novos.

O primeiro problema SEIR que consideramos tem um custo L^2 e restrições do estado puras. O nosso segundo problema tem um custo L^1 e restrições mistas. As soluções computacionais dos dois problemas são validadas parcialmente pelo princípio de máximo.

Para essa validação, normalidade e condições de regularidade são fundamentais. Ao resolver estes problemas não estamos a tratar uma doença particular numa população específica. Pretendemos assim ilustrar como formulações diferentes de controlo óptimo podem ser usadas para sugerir novas políticas de vacinação, tendo em consideração vários cenários e funcionais de custos diferentes.

O tratamento analítico da solução do problema L^2 com restrições do estado motiva o estudo de penalização exata para problemas com tais restrições. Definimos um conjunto de hipótese, segundo a qual a penalização exata escolhida assegura a regularidade dos multiplicadores.

A seguir, derivamos condições necessárias para problemas envolvendo equações diferenciais algébricas. Tal derivação baseia-se na aplicação de resultados conhecidos envolvendo problemas não suaves com restrições mistas. Notavelmente, não precisamos de aplicar o teorema da função implícita e tratamos alguns problemas com dados não suaves.

Finalmente, caracterizamos problemas de controlo envolvendo restrições do estado mistas usando inclusões diferenciais. Estabelecemos as propriedades das multifunções que definem a inclusão diferencial que nos permitem utilizar resultados sobre inclusões diferenciais já conhecidos na literatura.

Contents

Notation

$x + \varepsilon B$	Ball of radius ε centered at x in Euclidean space
B	Closed unit ball in Euclidean space
$[0, \infty]$	The range interval $[0, \infty)$ of a function while the value $\{+\infty\}$ may be assumed
$[-\infty, \infty]$	The range interval $(-\infty, \infty)$ of a function while the values $\{-\infty\}, \{+\infty\}$ may be assumed
$d_C(x)$	Euclidean distance of x to the set C
$\lvert \cdot \rvert$	Euclidean norm in $\mathbb{R}$
$\lVert \cdot \rVert_X$	Norm in the space X
$\lVert \cdot \rVert_1$	The norm of $L^1([a, b]; \mathbb{R}^p)$
$\lVert \cdot \rVert_\infty$	The norm of $L^\infty([a, b]; \mathbb{R}^p)$
$\lVert \mu \rVert_{TV}$	The norm of $C^*([a, b]; \mathbb{R})$
$N_A(x^*)$	Normal cone to a set A at the point x^*
$N_A^L(x^*)$	Limiting normal cone (also known as Mordukhovich normal cone) to A at x^*
$N_A^C(x^*)$	Clarke normal cone to a set A at the point x^*
$\partial^L f(x^*)$	Limiting subdifferential of a function f at the point x^*
$\partial^C f(x^*)$	Clarke subdifferential of a function f at the point x^*
$\bar{\partial}_x h$	The subdifferential of a function h with respect to state variable x
$\partial_x^> h$	The hybrid subdifferential of a function h with respect to state variable x
co $f(x^*)$	Convex hull of a function f at the point x^*

$\operatorname{supp}\mu$	Support of a measure μ
$C([a,b];\mathbb{R})$	Space of continuous functions from $[a,b]$ to $\mathbb{R}$
$C^*([a,b];\mathbb{R})$	Dual space of the space of continuous functions $C([a,b];\mathbb{R})$
$C^{\oplus}([a,b];\mathbb{R})$	Set of nonnegative elements in $C^*([a,b];\mathbb{R})$ on nonnegative valued functions in $C([a,b];\mathbb{R})$
$W^{1,1}([a,b];\mathbb{R})$	Space of absolutely continuous functions from $[a,b]$ to $\mathbb{R}$ (also denoted by $AC([a,b];\mathbb{R})$
$BV([a,b];\mathbb{R})$	Space of functions of bounded variation from $[a,b]$ to $\mathbb{R}$
$NBV([a,b];\mathbb{R})$	Space of functions of normalized bounded variation from $[a,b]$ to $\mathbb{R}$
$L^1([a,b];\mathbb{R}^n)$	Space of integrable (or L^1) functions from $[a,b]$ to $\mathbb{R}^n$
$L^\infty([a,b];\mathbb{R}^n)$	Space of essentially bounded (or L^∞) functions from $[a,b]$ to $\mathbb{R}^n$
$C^{1,1}$	Class of continuously differentiable functions with locally Lipschitz continuous derivatives
(x^*,u^*)	Optimal solution over all admissible processes for an optimal control problem
$\operatorname{Gr} f$	The graph of a function (or multifunction) f
$\operatorname{epi} f$	The epigraph of a function (or multifunction) f
$\operatorname{dom} f$	The domain of a function (or multifunction) f
$\operatorname{bdy} S$	The boundary of a set S
$\operatorname{cl} S$	The closure (also denoted by $\overline{S}$) of a set S
$\operatorname{int} S$	The interior of a set S

Chapter 1

Introduction

1.1 Motivation

Optimal control is where various areas of mathematics such as functional analysis, calculus of variations and nonsmooth analysis intertwine. Historically, optimal control has spun off from the calculus of variation and thus the historical perspective is a frequently chosen one for introduction. The survey [66] (and the references therein) gives an historic account of these pioneering results. The U.S. group of Hestenes, Isaacs and Bellman established a link to the calculus of variations, derived the necessary conditions and later developed the *dynamic programming principle*. The works of Pontryagin, Boltyanskii, Gamkrelidze, Mishchenko and others in the USSR led to the formulation of necessary conditions of optimality in the form of a *maximum principle*.

Variational systems are descriptive systems whereas optimal control problems differentiate between *control* and *state variables*, admit *control constraints*, and ultimately aim at steering the underlying dynamic systems into a desired direction via a *controlled differential equation* and achieving an *optimal solution*, i.e. minimizing costs or maximizing gain (more generally, a *value function*).

Many optimal control problems represent the real life behaviour of technical, biological or economical systems. Thus the limited physical or economical resources necessitate the use of constraints imposed on state or control variables. *Necessary conditions of optimality* must be formulated to account for these constraints.

1.2 Structure

The thesis is divided into two parts. **Part I** consists of three chapters and contains the fundamental concepts which we will explore in the second part of the thesis. **Chapter 2** begins with the general formulation of an optimal control problem, introduces state constraints and characterizes the different types of minimizers. Before proceeding with the necessary conditions of optimality we sidestep to define key elements of nonsmooth analysis, such as the *generalized gradient*, referring to the works of Clarke and others [13, 14, 75, 59, 15, 18]. This is necessary since real life control problems often have nonsmooth value functions and thus conventional, "smooth" necessary conditions fail to work in such cases. We then state necessary conditions for the case of smooth and the nonsmooth control problems. With regard to nonsmoothness and the presence of state constraints the presentation appeals to [20]. This chapter also highlights the fact that the adjoint multiplier associated with the state constraint is, roughly speaking, the "derivative of a function of bounded variation".

Chapter 3 then presents elements of measure theory which explain how the function of bounded variation, as beforementioned, associated with the ajoint multiplier of the state constraint, corresponds uniquely to a regular Borel measure. Of special interest is the decomposition of a regular unique Borel measure into a discrete, a singular and an absolutely continuous measure and the corresponding decomposition of a function of bounded variation. This theory is presented in a brief and self-contained manner and becomes relevant later when we investigate the regularity of the minimizer of the state constrained problem presented in Chapter 5.

Chapter 4 describes the compartmental SEIR control problem (named after the compartments "Susceptible", "Exposed", "Infectious" and "Recovered") which was proposed in [61]. This problem contrasts with other health related problems (see, for example, [29] and [72]). However a generic problem, the SEIR problem is used in praxis as a framework for modelling epidemic diseases with a more specific dynamics. The problem may utilize an L^2 or an L^1 cost which will be addressed in Chapters 5 and 7, respectively.

Part II is structured into five chapters to cover the area of conducted research. **Chapter 5** introduces a pure state constraint to the SEIR problem with an L^2 cost. To our knowledge, the introduction of state constraints to such problems has not occured until recently in [9], which is also the approach we adopt here. It turns out to be an apropriate testground for calculating and verifying analytic and numerical solutions for state-constrained problems with an L^2 cost.

Chapter 6 continues to focus on the L^2 state-constrained SEIR problem of Chapter 5. It presents an idea to overcome the discovered shortcoming in the study of the regularity of the measure ν at $t = T$ by creating an equivalent *penalized* problem without the state constraint, however, with an

additional penalization cost. The study of exact penalization is designed to provide a set of necessary conditions to determine an absolutely continuous ν.

What the remaining three chapters have in common is that the problems they examine can be treated as mixed control problems. These three chapters 7, 8 and 9 appeal to the nonsmooth maximum principle and the different *constraint qualifications* for mixed-constrained problems presented in [20].

More specifically, in **Chapter 7** we again analyse the SEIR problem, however, and with a mixed state constraint and with a cost functional of L^1 type, linear with respect to the control variable. It is argued that an L^1 type cost is more appropriate for biomedical control problems than an L^2 type cost.

Chapter 8 first classifies optimal control systems involving *differential-algebraic* equations (DAE). Applying necessary conditions and constraint qualifications from [20] we formulate new first order necessary conditions for DAE control systems systems, separately for the nonsmooth and smooth case. Hereby the algebraic variable can be treated as a control or as a state.

In **Chapter 9**, given a control problem (C) in terms of "conventional" functional equations, we investigate a corresponding problem in terms of differential inclusions, defined by a multifunction F, following the approach of [75], Chapter 2. If (C) contains a mixed-state constraint, we modify F to include the constraint already in its definition and call it F^-. The resulting control problem (DI), based on the differential inclusion $\dot{x}(t) \in F^-(t, x(t))$ a.e., is the focus of this chapter.

1.3 Contributions

Chapter 5 presents in a systematic way the analysis for validation of numerical solutions of a particular class of problems. In [9], although numerical results are compared with other problems, no theoretical discussion is made for the state constrained case and consequently no validation of the numerical solution is discussed. We partially remedy this in this thesis. Regarding the measure ν arising with adjoint multiplier associated with the state constraint we assert that the measure is absolutely continuous over the entire interval $[0, T)$. Although we are able to analytically verify regularity of ν with the theory of [73] in the interval $[0, T)$ this could not be affirmed for $t = T$. The analytically obtained results are compared against the numerically obtained ones via IPOPT/ICLOCS. The computed solution shows that in accordance with our findings (which verified the regularity of ν for the entire interval except $t = T$) the measure does exibit a jump at the end point. These findings were presented at *SADCO Summer School 2013* in Bayreuth. The case with both pure and mixed state constraints was presented at *FGP 2013*, Krakow and also published as [48].

In **Chapter 6**, we develop a set of necessary conditions for an equivalent penalized problem to the the state-constrained L^2 case of SEIR to ensure the absolute continuity of measure ν. We develop an additional hypothesis (HH) which, if asserted, ensures the validity of this approach. However, (HH) which essentially requires that, for a given admissible process to the original (constrained) problem there exists an admissible process to the penalized problem and the two trajectories are "close enough", is difficult to verify. These results were presented at *MTNS 2014*, Groningen and at SADCO workshop *New Perspectives in Optimal Control and Games*, 2014, Rome.

In **Chapter 7**, the L^1 type cost allows to define a *switching function* and so perform an analysis of *singular* vs. *bang-bang* optimal controls deduce both in closed form. The numerical solution is once more obtained via IPOPT/ICLOCS and corresponds exaclty to the analytical findings. These results were presented at *Controlo 2014*, Porto and are published as [27].

In **Chapter 8** we treat the algebraic equation as a mixed state constraint, apply the nonsmooth maximum principle in the wake of [20] and so formulate new first order necessary conditions for such a differential-algebraic system. Hereby we first treat the algebraic variable as a control. By introducing a differentiability assumption on the state constraint we are able to additionally simplify the Euler adjoint inclusion, i.e. to do without elements of the normal cone to the boundary of constraint. Notably, this approach gets by with no application of the implicit function theorem as it is otherwise often the case with DAE problems. Alternatively, we also present nonsmooth and smooth versions of the necessary conditions when the algebraic variable is seen as another "hybrid" state. This work was presented at *SADCO Doctoral Days 2012*, Paris, and also presented and published in the proceedings of *MTNS 2012*, Melbourne, [46].

After establishing the differential inclusion problem (DI) in **Chapter 9**, we investigate when the set of admissible trajectories of problem (C) and the set of admissible trajectores for (DI), are equal. We prove that the equality holds under a number of assumptions, most central the Lipschitz properties of the functions defining (C) and a bounded slope condition applying to the boundary of the admissible state-control set. The convexity of $F^-(t,x)$ is not one of the assumptions, however, assuming convexity we show that the set of F^- trajectories is compact with respect to the supremum norm topology and thus the problem (DI) has an optimal solution. Therefore we investigate and establish the logical implications between the convexity of $F^-(t,x)$, of $F(t,x)$ and other multifunctions involved. Via reformulation of the initial problem (C) into (DI), and asserting the previous assumptions we prove the existence of a minimizer for the originating problem (C). This work is submitted as [26] to *Set-Valued and Variational Analysis*.

1.4 Acknowledgements

At this place I would like to express my deep gratitude to my doctoral supervisor, Maria do Rosário de Pinho, for her continuous and persistent support. Her encouragement and inspiration during my last three years as her PhD student were always truly exceptional. She is a remarkable teacher and a wonderful person.

I am also very indebted to my co-supervisor Sofia Lopes for her support and invaluable suggestions. I would like to thank Matthias Gerdts for his extensive guidance and for a very enjoyable stay on my secondment at Bundeswehruniversität München. I am grateful for the fruitful advice of and the constructive talks with Margarida Ferreira and Fernando Fontes.

I thank my labmates, both former and present: Amélia Caldeira, Ana Filipa Ribeiro, Filipa Nogueira, Haider Biswas, Ian Marsh, Juliana Almeida, Luís Roque, Mario Lópes, Pedro Silva, Rui Calado and Tiago Paiva for an always cheerful atmosphere.

As a member of the European ITN SADCO project (Sensitivity Analysis for Deterministic Controller Design) which enabled my research position at University of Porto, I am obligated to its many organizers for letting me participate in its wide network of researchers in the related field. (Financial support was provided by the European Commission, FP7-PEOPLE-2010-ITN, Grant Agreement no. 264735-SADCO and is gratefully acknowledged. My work was also supported through the Portuguese Foundation for Science and Technology (FCT), within projects FCOMP-01-0124-FEDER-028894 and OCHERA PTDC/EEI-AUT/1450/2012.)

I am thankful to Lars Grüne for initially drawing my attention to the SADCO network.

I also thank my wife Jessica for her patience, and love.

Part I

Preliminary Concepts

Chapter 2

Basic Concepts of Optimal Control

We begin with a review of the basic notation which will be used throughout the latter chapters. An inequality $g(x) \leq 0$ in finite dimensional metric spaces, $g : \mathbb{R}^n \to \mathbb{R}^m$, is interpreted componentwise for $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Regardless the dimension of the underlying space, B is the unit ball centered at the origin and $|\cdot|$ is the Euclidean norm (in case of $\mathbb{R}^{p\times q}$, $p, q \in \mathbb{N}$, the induced matrix norm.) The *Euclidean distance function* to a set $A \subset \mathbb{R}^n$ is defined as $d_A : \mathbb{R}^n \to \mathbb{R}$,

$$d_A(x) := \inf\{|x - x'| : \ x' \in A\}.$$

The *signed distance function* $\tilde{d}_A : \mathbb{R}^n \to \mathbb{R}$ is defined as

$$\tilde{d}_A(x) := \begin{cases} d_A(x) & \text{if} \quad x \notin A \\ -d_{A^c}(x) & \text{if} \quad x \in A \end{cases} \tag{2.1}$$

where $A^c = \mathbb{R}^n \setminus A$ is the complement of A.

We will frequently make use of the following functional spaces. The spaces $L^1([a, b]; \mathbb{R}^n)$ and $L^\infty([a, b]; \mathbb{R}^n)$ are the spaces of *integrable* and *essentially bounded* functions, respectively. The space $C^*([a, b]; \mathbb{R})$ is the topological dual of the space of continuous functions $C([a, b]; \mathbb{R})$. Elements of $C^*([a, b]; \mathbb{R})$ can be identified with finite regular measures on the Borel subsets of $[a, b]$, as later explained in Chapter 3. The set of elements in $C^*([a, b]; \mathbb{R})$ taking nonnegative values on nonnegative-valued functions in $C([a, b]; \mathbb{R})$ is denoted by $C^\oplus([a, b]; \mathbb{R})$. The norm in $C^\oplus([a, b]; \mathbb{R})$, $|\mu|$, is equal with the total variation of μ, $\int_{[a,b]} \mu(ds)$. The support of a measure μ, written as $\text{supp}\{\mu\}$, is the smallest closed set $A \subset [a, b]$ such that for any relatively open subset $B \subset [a, b]\backslash A$ we have $\mu(B) = 0$.

2.1 Optimal Control Problems

There exist three common formulations of optimal control problems in a *fixed time*, namely the *Bolza form*, *Lagrange form*, and *Mayer form*. We begin with the statement of an optimal control problem of Bolza form and will later contrast it with the two other types. The problem comprises of a fixed interval $[a, b] \subset \mathbb{R}$, and the unknown variables $x = (x_1 \dots, x_n) \in \mathbb{R}^n$, called the *state variable*, and $u = (u_1 \dots, u_m) \in \mathbb{R}^m$, called the *control variable*.

Let the value of x be determined by an *absolutely continuous* function $x : [a, b] \to \mathbb{R}^n$, i.e. $x(t)$ is continuous and there exists a function $F \in L^1([a, b]; \mathbb{R}^n)$ such that[1]

$$x(t) = x(a) + \int_a^t F(s)\, ds, \quad s \in [a, b],$$

and the value of u be determined by a function $u : [a, b] \to \mathbb{R}^m$ which satisifies

$$u(t) \in U(t) \quad \text{a.e.}, \tag{2.2}$$

where U is a generally time-dependent closed set. The function u may be either measurable, continuous, integrable, piecewise continuous or, depending on the problem, defined otherwise.

Let a given *dynamics function* $f : [a, b] \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ define the *controlled* differential equation

$$\dot{x}(t) = f(t, x(t), u(t)) \quad \text{a.e. } t \in [a, b]. \tag{2.3}$$

If the before mentioned functions $x : [a, b] \to \mathbb{R}^n$ and $u : [a, b] \to \mathbb{R}^m$ solve (2.3) with a certain initial value $x(a) = x_0$, then x is called the *state trajectory* or the *state* function and u is called the *control* function. Together, the pair (x, u) is referred to as a *process*.

We call the set inclusion (2.2) the *control set constraint*. If a process (x, u) satisfies (2.2), (2.3) and besides, for a closed set $E \subset \mathbb{R}^n \times \mathbb{R}^n$, the *boundary condition*

$$(x(a), x(b)) \in E \tag{2.4}$$

then is called *admissible* for the following *optimal control problem*:

The problem is to determine the process (x^*, u^*) which minimizes a *cost functional* $J : W^{1,1}([a, b]; \mathbb{R}^n) \times L^1([a, b]; \mathbb{R}^m) \to \mathbb{R}$,

$$J(x, u) := l(x(a), x(b) + \int_a^b L(t, x(t), u(t))\, dt \tag{2.5}$$

[1] Another (equivalent) definition of an absolutely continuous function is Definition 3.3.12.

subject to the (already mentioned) conditions

$$\begin{aligned} &\dot{x}(t) = f(t, x(t), u(t)), && \text{a.e. } t \in [a, b], \\ &u(t) \in U(t), && \text{a.e. } t \in [a, b], \\ &(x(a), x(b)) \in E. \end{aligned}$$

over all admissible processes (x, u).

The minimizing process (x^*, u^*) is called an *optimal solution* to the optimal control problem (there may exist more than only one optimal solution). Furthermore, we may have locally or globally optimal solutions and speak, accordingly of *local* and *global minimizers*.

In summary, a optimal control problem in *Bolza form* is given by

$$(P_B) \quad \begin{cases} \text{Minimize } l(x(a), x(b)) + \displaystyle\int_a^b L(t, x(t), u(t))\, dt \\ \text{subject to} \\ \quad \dot{x}(t) = f(t, x(t), u(t)) \quad \text{a.e. } t \in [a, b] \\ \quad u(t) \in U(t) \quad \text{a.e. } t \in [a, b] \\ \quad (x(a), x(b)) \in E. \end{cases}$$

We choose now another cost functional which differs from (2.5) only by the missing term $l(x(a), x(b))$,

$$J(x, u) := \int_a^b L(t, x(t), u(t))\, dt. \tag{2.6}$$

Assuming the same dynamics function, control set constraint and the boundary condition, the resulting optimal control problem

$$(P_L) \quad \begin{cases} \text{Minimize } \displaystyle\int_a^b L(t, x(t), u(t))\, dt \\ \text{subject to} \\ \quad \dot{x}(t) = f(t, x(t), u(t)) \quad \text{a.e. } t \in [a, b] \\ \quad u(t) \in U(t) \quad \text{a.e. } t \in [a, b] \\ \quad (x(a), x(b)) \in E. \end{cases}$$

is said to be in *Lagrange form*. Finally, setting the cost functional

$$J(x, u) := l(x(a), x(b)) \tag{2.7}$$

defines an optimal control problem in *Mayer form*,

$$(P_M) \quad \begin{cases} \text{Minimize } l(x(a), x(b)) \\ \text{subject to} \\ \dot{x}(t) = f(t, x(t), u(t)) \quad \text{a.e. } t \in [a, b] \\ u(t) \in U(t) \quad \text{a.e. } t \in [a, b] \\ (x(a), x(b)) \in E. \end{cases}$$

The three forms are equivalent to each other in the sense that a given control problem of one of the three forms can be stated equivalently in one of the other two forms. The techniques for transformation between the three forms are well known and described e.g. in [5, 11, 22, 51, 52].

2.2 Optimal Control Problems with Constraints

As optimal control problems are often motivated by real life applications, the limited physical or economical resources may necessitate the use of state constraints. The limitations imposed on physical or economical resources imply many times *pathwise state constraints* on one or multiple state variables in form of *functional inequality constraints*. Such problems are called simply *state constrained optimal control problems*. However, there may also exist other types of of state constraints, such as *implicit* or *set-valued* constraints. If state constraints are present, they must be included into the necessary conditions for optimality.

Besides, we like to single out the *endpoint constraints* of a state trajectory. These constraints are so ubiquitous such they were already part of the three different problem forms formulated in (P_L), (P_B) and (P_M). Also the control set constraint mentioned in (2.2) may represent the limited access to a physical ressource in physical or engineering applications. In the following we want to categorize the above mentioned with a little more rigor.

Control Constraints

The statement $u(t) \in U(t)$ for almost every $t \in [a, b]$ is called the *control constraint*. In case $U \not\equiv const$, we require the multifunction $U : [a, b] \to \mathbb{R}^m$ to be measurable.

Endpoint Constraints

The *endpoint constraints* can be imposed with control problems over a fixed time interval $[a, b]$, which is the natural setting in the present work. The most general form of an endpoint constraint was stated in (2.4), i.e.

$$(x(a), x(b)) \in E \tag{2.8}$$

where E is a closed set. However, the following special cases are encountered frequently. Suppose the endpoint constraint reads as

$$\begin{cases} x(a) = x_a, \\ x(b) \in \mathbb{R}^n, \end{cases}$$

then writing $E = \{x_a\} \times \mathbb{R}^n$ makes the constraint equivalent to (2.8). Similarly, if E_b is a nonempty closed set,

$$\begin{cases} x(a) = x_a, \\ x(b) \in E_b, \end{cases}$$

translates to (2.8) with $E = \{x_a\} \times E_b$. Of course, it can be $E_b = \{x_b\}$, in such case the condition

$$\begin{cases} x(a) = x_a, \\ x(b) = x_b, \end{cases}$$

is translated to (2.8) by $E = \{x_a\} \times \{x_b\}$.

Also the initial and terminal state may be given in the form of functional equalities and/or inequalities. For example, we can have

$$E = \{(x, y) : \ \varphi_i(x(a), x(b)) \leq 0, \gamma_j(x(a), x(b)) = 0, \ i = 1, \ldots, k, \ j = 1, \ldots, l\}.$$

Pathwise Constraints

The most general way to describe the limited range of values which both the state and control variables can assume over the time interval $[a, b]$ (or any nonempty subinterval of it) is

$$(x(t), u(t)) \in C(t) \quad \text{for a.e. } t \in [a, b] \tag{2.9}$$

where $C : [a, b] \to \mathbb{R}^n \times \mathbb{R}^m$ is a given multifunction. However, one can find the following constraint forms:

Pure state constraints: For a function $h : [a,b] \times \mathbb{R}^n \to \mathbb{R}^k$ the condition

$$h(t, x(t)) \leq 0 \quad \text{for all } t \in [a,b] \tag{2.10}$$

is a *pure state constraint*. If it is clear from the context that the function h does not depend on the control u one may simply speak of *state constraints*. If the condition is formulated with the equality sign, to highlight this fact one speaks of *equality state constraints*.

Mixed state constraints: For a function $g : [a,b] \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^k$ the condition

$$g(t, x(t), u(t)) \leq 0 \quad \text{for a.e. } t \in [a,b] \tag{2.11}$$

is a *mixed state constraint*. Sometimes the name *mixed state-control constraint* is used to point out the depedency on both the state trajectory and the time-dependent control function. Also in this case the strict equality "=" may appear instead of "≤".

Note that pure state constraints are imposed *for all* t in an interval $[a,b]$ while mixed state constraints have to hold only *for almost every* $t \in [a,b]$. The third and more general type of constraints are

Implicit state constraints. We state versions for a pure and for a mixed state constraint. For given multifunctions $X : [a,b] \to \mathbb{R}^n, \tilde{X} : [a,b] \to \mathbb{R}^n \times \mathbb{R}^m$ the conditions

$$\begin{array}{rll} x(t) & \in X(t) & \text{for all } t \in [a,b], \quad \text{respectively,} \\ (x(t), u(t)) & \in \tilde{X}(t) & \text{for a.e. } t \in [a,b] \end{array} \tag{2.12}$$

are called *implicit state constraints*.

It is helpful to keep in mind that the (explicit) state or mixed state constraints can be transfered into implicit state constraints and vice versa using the following techniques:

Given $h(t, x(t)) \leq 0$ for all $t \in [a,b]$ as in (2.10) or $g(t, x(t), u(t)) \leq 0$ for a.e. $t \in [a,b]$ as in (2.11) we set

$$\begin{array}{rll} X(t) & := \{x \in \mathbb{R}^n : & h(t,x) \leq 0 \,\}, \quad \text{respectively,} \\ \tilde{X}(t) & := \{(x,u) \in \mathbb{R}^n \times \mathbb{R}^m : & g(t,x,u) \leq 0 \,\} \end{array}$$

and thus obtain the implicit constraint $x(t) \in X(t)$ for all $t \in [a,b]$ (resp., $(x(t), u(t)) \in \tilde{X}(t)$ for a.e. $t \in [a,b]$). If both pure and mixed state constraints are present, we may define an implicit state constraint

$$(x,u) \in (X(t) \times U(t)) \cap \tilde{X}$$

where X and $\tilde{X}$ are defined as above.

We show how to transform an implicit constraint into an explicit one if, for example, a pure state contraint $x(t) \in X(t)$ for all $t \in [a, b]$ is given. Then by setting $\tilde{h}(t, x(t)) := \tilde{d}_{X(t)}(x(t))$ with the signed distance $\tilde{d}$, defined in (2.1), we obtain

$$\tilde{d}_{X(t)}(x(t)) \leq 0 \quad \text{for all } t \in [a, b].$$

Returning to the state constraint $h(t, x(t)) \leq 0$ we say that, for a given trajectory x, the state constraint

(i) has a *boundary interval* if

$$\exists [t_0^b, t_1^b] : \quad h(t, x(t)) = 0 \quad \forall t \in [t_0^b, t_1^b],$$

in this case, t_0^b and t_1^b are called *entry* and *exit points*, respectively;

(ii) has a *contact point* $\sigma \in [a, b]$ if

$$h(\sigma, x(\sigma)) = 0 \quad \text{and} \quad \forall t \in [\sigma - \varepsilon, \sigma) \cup (\sigma, \sigma + \varepsilon] : \; h(x(t)) < 0;$$

(iii) has an *interior interval* if

$$\exists \, [t_0^i, t_1^i] : \quad h(t, x(t)) < 0 \quad \forall t \in (t_0^i, t_1^i).$$

Note that in literature one sometimes also distinguishes a *touch point* at time σ if it is a contact point and, additionally, $\frac{d}{dt} h(t, x(t))$ is continuous at σ (see, for example, [40]).

2.3 Minimizers

There exist two large classes of minimizers: *global* and *local minimizers*. Suppose for example, we want to find the minimizers of the problem

$$\text{Minimize } f(x) \text{ subject to } x \in \mathbb{R}^n. \tag{2.13}$$

Then x_G^* will be a *global minimizer* of (2.13), if it minimizes the cost over all other $x \in \mathbb{R}^n$, i.e.

$$f(x_G^*) \leq f(x) \;\; \forall x \in \mathbb{R}^n,$$

and x_L^* will be a *local minimizer* of (2.13), if it minimizes the cost over all other x in some neighbourhood, i.e. there exists $\varepsilon > 0$ such that

$$f(x_L^*) \leq f(x) \quad \forall x \in B(x_L^*; \varepsilon).$$

It was shown in [41] that in some cases the *local minimizers* are *global minimizers*. Furthermore, all global minimizers are local minimizers. We refer to [75, 77] for a study on minimizers. Throughout this thesis, we will restrict our discussion to minimizers in the context of optimal control problems to those of local minimizers.

Strong Local Minimizer

Suppose that (x^*, u^*) is an admissible process to a given optimal control problem with cost functional J. It is a *strong local minimizer* for an optimal control problem if, for some $\varepsilon > 0$, it minimizes the cost J, i.e.

$$J(x^*, u^*) \leq J(x, u)$$

for all other admissible processes (x, u) which satisfy

$$\| x - x^* \|_\infty \leq \varepsilon.$$

Remark: We have

$$\| x \|_\infty := \operatorname*{esssup}_{t \in [a,b]} |x(t)|,$$

so if $\|x\|_\infty \leq \varepsilon$ then $|x(t)| \leq \varepsilon$ for almost every $t \in [a, b]$. But if x is continuous, then

$$\operatorname*{esssup}_{t \in [a,b]} |x(t)| = \max |x(t)|.$$

So

$$|x(t) - x^*(t)| \leq \varepsilon \quad \text{for all } t \in [a, b] \quad \Longleftrightarrow \quad \max |x(t) - x^*(t)| \leq \varepsilon.$$

Weak Local Minimizer

Suppose again that (x^*, u^*) is an admissible process. It is called a *weak minimizer* if there exists $\epsilon > 0$ such that

$$J(x^*, u^*) \leq J(x, u)$$

holds for all processes (x, u) which satisfy

$$|x(t) - x^*(t)| \leq \varepsilon \quad \forall t \in [a, b] \qquad \text{and} \qquad |u(t) - u^*(t)| \leq \varepsilon \quad \text{for a.e. } t \in [a, b].$$

$W^{1,1}$ Local Minimizer

The process (x^*, u^*) is a $W^{1,1}$ *local minimizer* for an optimal control problem if, for some $\varepsilon > 0$, it minimizes the cost over all other admissible processes (x, u) such that

$$\| x - x^* \|_\infty \leq \varepsilon, \quad \text{and} \quad \int_a^b |\dot{x}(t) - \dot{x}^*(t)| dt \leq \varepsilon.$$

Recall that

$$\begin{aligned} \| x - x^* \|_{W^{1,1}} &= |x(a) - x^*(a)| + \| \dot{x}(t) - \dot{x}^*(t) \|_{L^1} \\ &= |x(a) - x^*(a)| + \int_a^b |\dot{x}(t) - \dot{x}^*(t)| dt. \end{aligned}$$

Local Minimizer of Radius R

Let us define a measurable function $R : [a, b] \to (0, +\infty]$ which is called a *radius function*. The process (x^*, u^*) is a *local minimizer of radius* R for an optimal control problem if, for some $\varepsilon > 0$, it minimizes the cost over all other admissible processes (x, u) satisfying

$$\| x - x^* \|_\infty \leq \varepsilon \quad , \quad \int_a^b |\dot{x}(t) - \dot{x}^*(t)| dt \leq \varepsilon,$$

as well as

$$|u(t) - u^*(t)| \leq R(t), \quad \text{a.e.} \quad t \in [a, b].$$

Remark: The relationship between strong and weak local minimizers is that a strong local minimizer is always a weak local minimizer but the converse is not necessarily true [77]. Throughout this thesis we mostly work with *strong minimizers*.

2.4 Elements of Nonsmooth Analysis

Nonsmooth analysis comes into place when classical analysis fails to provide approximation to non-differentiable functions or sets of functions with nondifferentiable boundaries. The theory has been pioneered in the 1970s by F. H. Clarke generalizing the concept of the subdifferential of a convex function. The so-called *generalized gradients* allow to formulate necessary conditions for optimality for nonsmooth problems. The application of nonsmooth analysis to optimal control since then has been an area of research and is highlighted, for instance, in [13], [14] and [75].

In the *classical* sense, derivatives of a function f are related to normal vectors to tangent hyperplanes; for any differentiable function f the vector $(f'(x), -1)$ is a downward normal to the graph of f at $(x, f(x))$. The graph of f is defined by $\mathrm{Gr}f = \{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} : \alpha = f(x)\}$. This geometric relationship is the key for the development of nonsmooth analysis. Instead of considering derivatives as elements of normal subspaces to smooth sets, *generalized derivatives* are defined to be elements of normal cones to possibly nonsmooth sets.

Let $A \subset \mathbb{R}^n$ be a nonempty closed set with $x \in \mathbb{R}^n \backslash A$. We call y the *closest point* in A or $proj_A(x)$ (i.e. the *projection* of x onto A) (see Figure 2.1) if y is such that

$$\| x - y' \| \geq \| x - y \|, \ \forall y' \in A$$

which is equivalent to

$$\langle \omega, y' - y \rangle \leq \sigma \| y' - y \|^2, \ \forall y' \in A \ \text{ and some } \sigma > 0,$$

where the vector $\omega = x - y$ is orthogonal to A at y.

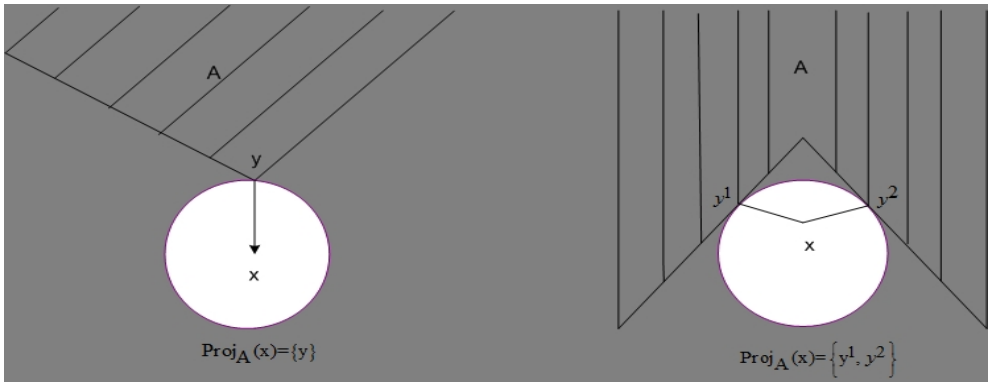


Figure 2.1: Geometrical interpretation of proximal normal and limiting normal cones.[2]

[2]Source: [28]

Any nonnegative multiple $\zeta = t\omega,\ t > 0$ of ω is called a *proximal normal* vector. That is, a *vector* ζ is called a *proximal normal* to A at y iff for some $\sigma > 0$ the following *proximal normal inequality* holds:

$$\langle \zeta, y' - y \rangle \leq \sigma \parallel y' - y \parallel^2,\ \forall y' \in A.$$

The set of all such vectors, which is a convex cone containing 0 is denoted by $N_A^P(y)$ and is called the *proximal normal cone*.

A vector ζ is called the *limiting normal to A at x* if for each $i \in \mathbb{N}$,

$$\zeta = \lim \zeta_i,\ \forall \zeta_i \in N_A^P(x_i),\ x_i \in A,\ x_i \to x,$$

and the set of all such limiting normals, denoted by $N_A^L(x)$ is a cone, called the *limiting normal cone* to A at x.

Given a lower semicontinuous function $f \colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and a point $x \in \mathbb{R}^n$ where $f(x) < +\infty$ such that $\mathrm{dom} f = \{x : f(x) < +\infty\}$, then the *proximal subdifferential* (or set of all *proximal subgradients*) of f at $x \in \mathrm{dom} f$ is defined as the set

$$\partial^P f(x) := \{\zeta \in \mathbb{R}^n : (\zeta, -1) \in N_{\mathrm{epi} f}^P(x, f(x))\}.$$

where $\mathrm{epi} f = \{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} : \alpha \geq f(x)\}$ denotes the epigraph of a function f. Alternatively, all $\varsigma \in \partial^P f(x)$ can be defined as those sufficing the condition

$$f(y) \geq f(x) + \langle \varsigma, y - x \rangle - \sigma \left| y - x \right|^2$$

for certain $\delta > 0, \sigma \geq 0$ and all $y \in x + \delta B$. The *limiting subdifferential* (or set of all *limiting subgradients*) of a function f at $x \in \mathrm{dom} f$ denoted by $\partial^L f(x)$ is obtained by the set

$$\partial^L f(x) := \{\zeta \in \mathbb{R}^n : (\zeta, -1) \in N_{\mathrm{epi} f}^L(x, f(x))\}.$$

Notably, the nonsmooth calculus can be developed via the theory of *generalized gradients* in the context of *locally Lipschitz function*. If a function $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz near x, then the *generalized gradient* $\partial^C f(x)$ can be written as

$$\partial^C f(x) = \mathrm{co}\, \partial^L f(x)$$

(*convex hull* of $\partial^L f(x)$); also similarly, the associated normal cone $N_A^C(x)$ to a set A at a point x is

given by

$$N_A^C(x) = \overline{\text{co}}\, N_A^L(x).$$

Since the generalized gradient and its calculus were first defined by Clarke in 1973 [13], $\partial^C f(x)$ and $N_A^C(x)$ are also called the *Clarke subdifferential* and the *Clarke normal cone* respectively. For more details on such nonsmooth analysis concepts and generalized gradients as well as its basic calculus, we refer e.g. to [13, 14, 59, 75].

2.5 The Maximum Principle

The *Maximum Principle* (MP) provides a set of necessary conditions which must be satisfied by any optimal solution to a given optimal control problem. There are different versions taylored to different type of control problems; a *smooth* maximum principle is used when the data of the problem are smooth, a *nonsmooth* maximum principle is required when the data are nonsmooth. We have a maximum principle for problems *with* and *without* state constraints.

The idea behind the maximum principle is to obtain necessary conditions describing the smallest set of possible solutions as possible. In some cases the maximum principle is not only a necessary condition for optimality but also a sufficient condition.

Smooth Maximum Principle

The smooth version of the maximum principle is the version of a maximum principle that the group of Soviet researchers led by Pontryagin came up with in the 1960s and the first one which became widely accepted. Assume a Mayer optimal control problem (without state constraints in the first place).

$$\text{(P)} \quad \begin{cases} \text{Minimize } l(x(a), x(b)) \\ \text{subject to} \\ \dot{x}(t) = f(t, x(t), u(t)) \quad \text{a.e. } t \in [a, b] \\ u(t) \in U(t) \quad \text{a.e. } t \in [a, b] \\ (x(a), x(b)) \in E. \end{cases}$$

We may assume that the data of the problem are sufficiently smooth, i.e., for example, that the functions f and l are continuously differentiable. Assume also, for simplicity, that the multifunction U does not depend on time (i.e., $(U(t) \equiv U)$ and U is a closed set. We define the *Pseudo-Hamiltonian*

(or *Unmaximized Hamiltonian*)

$$H(t,x,p,u) = \langle p, f(t,x,u)\rangle.$$

Now the *smooth maximum principle* for the problem (P) without state constraints under some appropriate assumptions can be presented in the next Theorem (an adaptation of Theorem 6.2.1 in [75]).

Theorem 2.5.1 (The Maximum Principle for (P) Without State Constraints) Let (x^*, u^*) be a strong local minimum for problem (P) without state constraints. Then there exist an arc $p \in W^{1,1}([a,b];\mathbb{R}^n)$ and a scalar $\lambda_0 \geq 0$ satisfying

the *Nontriviality Condition* [NT]:

$$(p, \lambda_0) \neq (0,0),$$

the *Euler Adjoint Equation* [AE]:

$$-\dot{p}(t) = \nabla_x \langle p(t), f(t, x^*(t), u^*(t))\rangle \quad \text{a.e.},$$

the global *Weierstrass Condition* [W]:

$$\forall \ \ u \in U,$$

$$\langle p(t), f(t,x^*(t),u)\rangle \leq \langle p(t), f(t,x^*(t),u^*(t))\rangle \quad \text{a.e.},$$

and the *Transversality Condition* [T]:

$$(p(a), -p(b)) = \lambda_0 \nabla l(x^*(a), x^*(b)) + (\eta_1, \eta_2),$$

for some $(\eta_1, \eta_2) \in N_E^L(x^*(a), x^*(b))$.

The function p is called the *costate* (or *adjoint*) function and λ_0 the *cost multiplier*. The adjoint equation is also called the *costate differential equation*.

We now turn our attention to the more general case of the problem (P), the problem ($\mathrm{P_S}$):

$$(\mathrm{P_S}) \quad \begin{cases} \text{Minimize } J(x,u) = l(x(a), x(b)) \\ \text{subject to} \\ \dot{x}(t) = f(t, x(t), u(t)) \quad \text{a.e. } t \in [a,b] \\ h(t, x(t)) \leq 0 \quad \text{ for all } \ t \in [a,b] \\ u(t) \in U(t) \quad \text{ a.e. } \ t \in [a,b] \\ (x(a), x(b)) \in E, \end{cases}$$

which differs from (P) by the presence of the pathwise state constraint $h(t, x(t)) \leq 0$ imposed via

a function $h : [a,b] \times \mathbb{R}^n \to \mathbb{R}$. The effect of the additional state constraint is the appearance of *measures* as multipliers. The adjoint multiplier p will have to be replaced by a function q of *bounded variation* defined by

$$q(t) = \begin{cases} p(t) + \displaystyle\int_{[a,t)} \nabla h(s, x^*(s))\, \mu(ds), & t \in [a,b) \\ p(t) + \displaystyle\int_{[a,b]} \nabla h(s, x^*(s))\, \mu(ds), & t = b, \end{cases} \tag{2.14}$$

where $\mu \in C^{\oplus}([a,b])$. Functions of bounded variation will be defined in Chapter 3.

Let us assume again that the functions f, l and h are all continuously differentiable and, as before, that U is a closed set. Then the *smooth maximum principles* for the *state constrained optimal control problems* can be adapted in the following Theorem (an adaptation of Theorem 9.3.1 in [75]).

Theorem 2.5.2 (The Maximum Principle for (P_S) With State Constraints) Let (x^*, u^*) be a strong local minimum for problem (P_S) with state constraints. Then there exists an arc $p \in W^{1,1}([a,b]; \mathbb{R}^n)$, an arc $q \in BV([a,b]; \mathbb{R}^n)$, a scalar $\lambda_0 \geq 0$ and $\mu \in C^{\oplus}([a,b])$, such that the following conditions are satisfied:

(i) The Nontriviality Condition [NT]:

$$(p, \mu, \lambda_0) \neq (0, 0, 0)$$

(ii) The Euler Adjoint Equation [AE]:

$$-\dot{p}(t) = \nabla_x \langle q(t), f(t, x^*(t), u^*(t)) \rangle \quad \text{a.e.},$$

(iii) The Weierstrass Condition [W]:

$$\forall \ u \in U,$$
$$\langle q(t), f(t, x^*(t), u) \rangle \leq \langle q(t), f(t, x^*(t), u^*(t)) \rangle \quad \text{a.e.},$$

(iv) The Transversality Condition [T]:

$$(p(a), -q(b)) = \lambda_0 \nabla l(x^*(a), x^*(b)) + (\eta_1, \eta_2),$$

for some $(\eta_1, \eta_2) \in N_E^L(x^*(a), x^*(b))$,

(v) : $\quad \text{supp}\{\mu\} \subset \{t \ : \ h(t, x^*(t)) = 0\}$.

where p and q are related by (2.14).

Remark 2.5.3

(i) Theorem 2.5.2 presents a version of the maximum principle for pure state constrained problems. We will discuss the mixed constrained case in Chapters 5 and 7.

(ii) Note also that the maximum principle in Theorem 2.5.2 is of interest only if the control problem is *normal*. As an illustration, a normality criterion is applied in Section 5.4.

Nonsmooth Maximum Principle

We now discuss here the more general *nonsmooth maximum principle* for optimal control problems with state constraints. In the 1970s Clarke generalized the convex subdifferentials of Rockafellar to cover Lipschitz continuous functions and, to some extent, lower semi-continuous functions (see, for example [13]). He also successfully applied nonsmooth analysis to optimization and optimal control theory. In 1976 Mordukhovich proposed the concept of limiting subdifferential and he showed how transversality conditions in the nonsmooth maximum principle could be weakened.

On a practical note, to exemplify the more general nature of the nonsmooth maximum principle we again call attention to the relationship between the adjoint arcs p and q which was already pointed out in (2.14) for the smooth case. Generally, the maximum principle is formulated with the relationship

$$q(t) = \begin{cases} p(t) + \displaystyle\int_{[a,t)} \gamma(s)\, d\nu(s), & t \in [a,b), \\ p(t) + \displaystyle\int_{[a,b]} \gamma(s)\, d\nu(s), & t = b, \end{cases} \tag{2.15}$$

where $\gamma : [a,b] \to \mathbb{R}^n$ is a measurable function satisfying

$$\gamma(t) \in \partial_x^{>} h(t, x^*(t)) \quad \mu\text{-a.e.}$$

with the *partial subdifferential* $\partial_x^{>} h(t,x)$ defined as

$$\partial_x^{>} h(t,x) := co\,\{\gamma : \exists (t_i, x_i) \xrightarrow{h} (t,x) : h(t_i,x_i) > 0\ \forall i,\ \nabla_x h(t_i,x_i) \to \gamma\} \tag{2.16}$$

(see [75] for reference). The function ν in (2.15) is of bounded variation such that $\nu(t)$ is constant on any interior interval, i.e. on $\{[t_0,t_1] \subset [a,b] \ :\ h(t,x(t)) < 0\ \forall\, t \in (t_0,t_1)\}$.

We will learn in Chapter 3 that, on one hand, functions of bounded variation are the dual space $C^*([a,b];\mathbb{R})$ and, on the other hand, every bounded variation function uniquely corresponds to a

certain Borel measure μ such that

$$\mu(I) = \int_I d\nu(t)$$

for all closed subintervals $I \subset [a, b]$. This allows us to conveniently re-write (2.15) as

$$q(t) = \begin{cases} p(t) + \displaystyle\int_{[a,t)} \gamma(s)\, \mu(ds), & t \in [a, b) \\ p(t) + \displaystyle\int_{[a,b]} \gamma(s)\, \mu(ds), & t = b. \end{cases} \tag{2.17}$$

while, somewhat casually, saying $\mu \in C^{\oplus}([a, b])$. On the other hand, we have in Theorem 2.5.2

$$\partial_x^{>} h(t, x^*(t)) = \nabla h(t, x^*(t))$$

since the function h continuously differentiable, and the formulation (2.17) is equivalent to the previously stated in (2.14).

In this context, another relevant fact is that any function ν of bounded variation can be decomposed uniquely as

$$\nu = \varphi + r + s,$$

where φ is an absolutely continuous function, r is a singular function and s is a jump function. In a similar way the unique Borel measure μ corresponding to ν can be written as

$$\mu = \mu_{ac} + \mu_{sc} + \mu_d,$$

where ac stands for absolutely continuous, sc for singular and d for discrete measure. The equivalence between measures and bounded variation functions is widely known and yet will be shown explicitly, for the sake of clarity, in the next Chapter 3.

One important consequence is the following one: the multiplier $q(t)$ may have jumps on the boundary of the state constraint. This is exactly the case when the discrete measure μ_d happens to be nonzero. In many applications, however, it is desirable to have absolutely continuous adjoint multipliers, thus, it is important to know whether jumps occur or not. Chapter 5 contains a practical study of a given problem and a numerical proof that jumps do occur in the case of the SEIR control problem, at least once, at the very end of the time interval $[0, T]$.

We consider again our problem (P_S) with state constraints, but we will impose the following hypotheses which make reference to an optimal solution (x^*, u^*) and a parameter $\varepsilon > 0$:

(NH1): The function $(t,u) \to f(t,x,u)$ is $\mathcal{L}\times\mathcal{B}$ measurable [3] and there exist $\varepsilon > 0$ and an integrable function $k(t)$ such that, for almost every $t \in [a,b]$ the following condition holds:

$$|f(t,x_1,u) - f(t,x_2,u)| \le k(t)\|x_2 - x_1\|, \ \forall u \in U(t), \ (x_1,x_2) \in B(x^*,\varepsilon).$$

(NH2): l is Lipschitz near $(x^*(a),x^*(b))$ with Lipschitz constant K_l.

(NH3): h is upper semicontinuous and for each $t \in [a,b]$ the function $x \mapsto h(t,x)$ is Lipschitz on $x^*(t) + B(0,\varepsilon)$ with Lipschitz constant K_h.

(NH4): $Gr\,U$ is a Borel set, where $Gr\,U$ is defined as

$$GrU := \{(t,u) \in [a,b] \times \mathbb{R}^m : u \in U(t)\},$$

the *graph* of a multifunction $U : [a,b] \to \mathbb{R}^m$.

Theorem 2.5.4 (The Nonsmooth Maximum Principle for (P_S) With State Constraints) Let (x^*,u^*) be a strong local minimum for problem (P_S) with state constraints and assume that hypotheses (NH1)–(NH4) are satisfied. Then there exist an arc $p \in W^{1,1}([a,b];\mathbb{R}^n)$, an arc $q \in BV([a,b];\mathbb{R}^n)$, a scalar $\lambda_0 \ge 0$, $\mu \in C^{\oplus}([a,b])$, and a measurable function $\gamma(t) : [a,b] \to \mathbb{R}^n$ satisfying $\gamma(t) \in \partial_x^{\ge} h(t,x^*(t)) \ \ \mu - a.e.$ such that the following conditions are satisfied.
(i) The Nontriviality Condition [NT]:

$$(p,\mu,\lambda_0) \ne (0,0,0),$$

(ii) The Euler Adjoint Equation [AE]:

$$-\dot{p}(t) \in \partial_x^C \langle q(t), f(t,x^*(t),u^*(t))\rangle \ \text{ a.e.},$$

(iii) The Weierstrass Condition [W]:
$\forall \ u \in U(t)$,

$$\langle q(t), f(t,x^*(t),u)\rangle \le \langle q(t), f(t,x^*(t),u^*(t))\rangle \ \text{ a.e.},$$

(iv) The Transversality Condition [T]:

$$(p(a),-q(b)) \in \lambda_0 \partial l(x^*(a),x^*(b)) + (\eta_1,\eta_2),$$

for some $(\eta_1,\eta_2) \in N_E^C(x^*(a),x^*(b))$,

[3] We anticipate here the definition of "Borel measurable" and "Lebesgue measurable" in Section 3.4.

(v) : $\quad \operatorname{supp}\{\mu\} \subset \{t \,:\, h(t, x^*(t)) = 0\},$

where q is as in (2.17), and the partial subdifferential $\partial_x^{>}$ is as defined in (2.16).

Remark 2.5.5 Several extended versions, and even more strengthened forms of the nonsmooth maximum principle for state constrained optimal control problems have been developed over the years. We refer to [13, 16, 17, 21, 74] for the detailed presentations and to [19, 20] for the recent developments in the nonsmooth maximum principle.

Chapter 3

Elements of Measure Theory

The aim of this chapter is to gather, in a (possibly) straightforward way, some theoretical results that will be of relevance in the following chapters. We focus our attention on the relation between measures, decomposable into a discrete, a singular and an absolutely continuous measure, and functions of bounded variation, decomposable into a jump function, a singular and an absolutely continuous function.

We do not intend to give an extensive introduction to measure theory. Indeed, many concepts are omitted as they are assumed to be known or not relevant to the future chapters. When proofs are given, it will be to illustrate the related concepts.

3.1 Measures

As a reminder, we begin with the definition of a σ-algebra. The Borel σ-algebra will be introduced later in this chapter. Let X denote a metric space or, more generally, a topological space. Let $\mathcal{P}(X)$ be the *power set* of X, i.e. the collection of all subsets of X.

Definition 3.1.1 A subset $\mathcal{M}$ of $\mathcal{P}(X)$ is called a *σ-algebra* if the following holds:

(i) $\mathcal{M}$ is nonempty;

(ii) If $E \in \mathcal{M}$, then $X \setminus E \in \mathcal{M}$;

(iii) The union of countably many sets in $\mathcal{M}$ is also in $\mathcal{M}$, i.e. if $E_i \in \mathcal{M}$ for $i \in \mathbb{N}$, then $\bigcup_{i=1}^{\infty} E_i \in \mathcal{M}$.

Definition 3.1.2 Let $\mathcal{E}$ be a collection of sets of X. The intersection of all σ-algebras containing $\mathcal{E}$ is called the *σ-algebra generated by $\mathcal{E}$*, denoted by $\sigma(\mathcal{E})$.

Definition of a measure and decomposition results

Definition 3.1.3 (Measure) Let $\mathcal{M}$ be a σ-algebra of $\mathbb{R}$. A *measure* on $(X, \mathcal{M})$ is a function $\nu : \mathcal{M} \to [0, \infty)$ such that the following properties are satisfied:

(i) $\nu(\emptyset) = 0$;

(ii) ν can assume the value of ∞;

(iii) *(Countable additivity)* If $\{E_j\}$ is a sequence of disjoint sets in $\mathcal{M}$, then

$$\nu\left(\bigcup_{j=1}^{\infty} E_j\right) = \sum_{j=1}^{\infty} \nu(E_j),$$

whereas $\sum_{j=1}^{\infty} \nu(E_j)$ converges absolutely if $\nu\left(\bigcup_{j=1}^{\infty} E_j\right)$ is finite.

Definition 3.1.4 (Signed measure) Let $\mathcal{M}$ be a σ-algebra of $\mathbb{R}$. A *signed measure* on $(X, \mathcal{M})$ is a function $\nu : \mathcal{M} \to (-\infty, \infty)$ such that the properties (i), (iii) of Definition 3.1.3 are satisfied and, instead of (ii), the following holds:

(ii) ν assumes at most one of the values $\pm\infty$.

Remark 3.1.5 $(X, \mathcal{M})$ is called a *measurable space*, and $(X, \mathcal{M}, \nu)$ is called a *measure space*.

Two special kinds of a measure will sometimes play an important role:

Definition 3.1.6 Let $\mathcal{M}$ be a σ-algebra of $\mathbb{R}$.

(i) μ is called the *counting measure* on $\mathcal{M}$ if

$$\mu(E) = \sum_{x \in E} 1 \quad \forall E \in \mathcal{M};$$

(ii) μ is called the *Dirac measure at x_0* if, for some $x_0 \in X$

$$\mu(E) = \begin{cases} 1, & x_0 \in E, \\ 0, & x_0 \notin E. \end{cases}$$

Definition 3.1.7 Let ν be a signed measure on $(X, \mathcal{M})$. A set $E \in \mathcal{M}$ is called *positive* (respectively, *negative*, *null*) if with respect to ν, if

$$\nu(F) \geq 0 \quad (\text{resp. } \nu(F) \leq 0,\ \nu(F) = 0) \quad \text{for all } F \in \mathcal{M} \text{ such that } F \subset E$$

Theorem 3.1.8 (Hahn decomposition theorem, e.g. [36] Thm. 3.3) Let ν be a signed measure on $(X, \mathcal{M})$. Then there exists a ν-positive set P and a ν-negative set N such that $P \cup N = X$, $P \cap N = \emptyset$.

The sets P, N are called the *Hahn decomposition of X with respect to ν*. The Hahn decomposition is not unique but if there are sets P', N' which satisfy the same criteria, then $\nu(P \Delta N) = 0$[1]. That is, the sets P, N and P', N' differ at most by a set S with $\nu(S) = 0$.

For the proof of the theorem see [36]. It is quite technical though interesting because it emphasizes the fact that the $\nu - positive$, $\nu - negative$ property of, respectively, sets P, N does not necessarily follow from $\nu(P) \geq 0,\ \nu(N) \leq 0$. It is rather required that every possible subset of P is of measure ≥ 0 and every possible subset of N is of measure ≤ 0.

Definition 3.1.9 Let μ, ν be two signed measures on $(X, \mathcal{M})$. The measure ν is *singular* with respect to μ (equivalently, μ is *singular* w.r.t. ν), denoted by

$$\mu \perp \nu,$$

if there exist $E, F \in \mathcal{M}$ such that

$$E \cap F = \emptyset, \quad E \cup F = X \quad \text{and} \quad \mu(E) = 0, \quad \nu(F) = 0,$$

in other words "μ and ν live on disjoint sets".

Definition 3.1.10 Let ν be a signed measure and μ a positive measure on $(X, \mathcal{M})$. The measure ν is *absolutely continuous* with respect to μ, denoted by

$$\nu \ll \mu,$$

if $\nu(E) = 0$ for every $E \in \mathcal{M}$ for which $\mu(E) = 0$.

Definition 3.1.10 can be alternatively described by the following theorem. This property of absolutely continuous measures will be later helpful to relate them to absolutely continuous functions.

[1] Δ denotes the *symmetric difference* of sets, i.e. $P \Delta N = (P \setminus N) \cup (N \setminus P)$.

Theorem 3.1.11 ([36] Thm. 3.5) Let ν be a finite signed measure and μ a positive measure on $(X, \mathcal{M})$. Then $\nu \ll \mu$ iff for every $\epsilon > 0$ there exists $\delta > 0$ such that $|\nu(E)| < \epsilon$ whenever $\mu(E) < \delta$.

The Hahn decomposition $P \cup N$ allows to construct the measure μ as the difference of two positive measures. This becomes visible in the proof of the Jordan decomposition theorem by the construction $\nu^+(E) = \nu(E \cap P)$ and $\nu^-(E) = \nu(E \cap N)$.

Theorem 3.1.12 (Jordan Decomposition, [36] Thm. 3.4) Let ν be a signed measure. Then there exist unique positive measures ν^+ and ν^- such that

$$\nu = \nu^+ - \nu^- \quad \text{and} \quad \nu^+ \perp \nu^-.$$

Definition 3.1.13 Let ν be a signed measure and ν^+, ν^- be defined as in Theorem 3.1.12.

(i) The measures ν^+ and ν^- are called the *positive* and the *negative* variation of ν;

(ii) The difference $\nu^+ - \nu^-$ is called the *Jordan decomposition* of ν;

(iii) The sum $\nu^+ + \nu^-$ is called the *total variation* of ν and denoted $|\nu|$.

Properties of signed measures

Let μ, ν be signed measures defined on $(X, \mathcal{M})$, $\mu, \nu : \mathcal{M} \to [-\infty, \infty]$.

1. $\forall\, E \in \mathcal{M}$ such that $\nu(E) = 0 \iff |\nu|(E) = 0$,
2. $\nu \perp \mu \iff |\nu| \perp \mu$,
3. if $\nu < \infty$ on $\mathcal{M}$, then $\nu^+(X) = \nu(P)$, where P is a ν-positive set, $P \in \mathcal{M}$,
4. if ν, μ are measures such that $\nu \ll \mu$ and $\nu \perp \mu$ then $\nu = 0$,
5. if ν, μ are measures then $\nu \ll \mu$ holds if and only if $|\nu| \ll \mu$,
6. if ν, μ are measures then $|\nu| \ll \mu$ holds if and only if $\nu^+ \ll \mu$, $\nu^- \ll \mu$,

3.2 Radon-Nikodym Theorem and the Lebesgue Decomposition

In what follows, the integral $\int$ will denote the standard Lebesgue integral. The space $L^1([a,b],m)$ is the space of functions which are Lebesgue-integrable almost everywhere on $[a,b]$, i.e.

$$f \in L^1([a,b],m) \qquad \Longleftrightarrow \qquad \int_a^b f\,dm < \infty,$$

where m denotes the Lebesgue measure[2].

Definition 3.2.1 Consider two measurable spaces $(X,\mathcal{M}_1),(Y,\mathcal{M}_2)$ and a function $f : X \to Y$. The function f is called $(\mathcal{M}_1,\mathcal{M}_2)$-*measurable* if

$$f^{-1}(E) = \{x \in X \,:\, f(x) \in E\} \in \mathcal{M}_1 \quad \forall\, E \in \mathcal{M}_2.$$

The function $f : \mathbb{R} \to \mathbb{R}$ is called *Borel-measurable*, if

$$f^{-1}(O) = \{x \in Dom(f) \,:\, f(x) \in O\} \in \mathcal{B}_{\mathbb{R}}$$

for any open set O, where $\mathcal{B}_m r$ is the Borel σ-algebra[3].

Theorem 3.2.2 (Radon-Nikodym) Let ν be a σ-finite signed measure and μ a σ-finite measure on $(X,\mathcal{M})$ such that

$$\nu \ll \mu$$

Then there exists a measurable function $f : X \to \mathbb{R}$ such that at least one of the integrals $\int f^+\,d\mu$ or $\int f^-\,d\mu$ is finite , where f^+, f^- are such that $f = f^+ - f^-$ holds, and

$$\nu(E) = \int_E f\,d\mu \quad \forall\, E \in \mathcal{M}.$$

If there exists another function g such that $\nu(E) = \int_E g\,d\mu \quad \forall\, E \in \mathcal{M}$, then $f = g\ \mu - a.e.$

Theorem 3.2.3 (Lebesgue-Radon-Nikodym (also Lebesgue decomposition)) Let ν be a σ-finite signed measure and μ a σ-positive measure on $(X,\mathcal{M})$. There exists unique σ-finite signed

[2] See definition of the Lebesgue measure in Remark 3.4.4.
[3] See Definition 3.4.1

measures λ, ρ on $(X, \mathcal{M})$ such that

$$\lambda \perp \mu, \quad \rho \ll \mu, \quad \text{and} \quad \nu = \lambda + \rho,$$

and the measures ρ and λ are unique.

Remark 3.2.4 The above decomposition,

$$\nu = \lambda + \rho, \text{where} \quad \lambda \perp \mu \quad \text{and} \quad \rho \ll \mu,$$

is called the *Lebesgue decomposition of ν with respect to μ*. Any σ-finite signed measure ν defined in a measure space $(X, \mathcal{M}, \mu)$ can be decomposed into an absolutely continuous measure with respect to μ and a singular measure with respect to μ. Note that if either ν or μ is not σ-finite then the Lebesgue decomposition may fail. Consider Theorem 3.2.5 (Exercise 13 in [36], p. 92).

In the case of $\nu \ll \mu$ Theorem 3.2.3 yields that

$$d\nu = f\, d\mu \quad \text{for some } f.$$

This is also the direct result of Theorem 3.2.2. The function f in this case is known as the *Radon-Nikodym derivative of ν with respect to μ*, denoted as $\frac{d\nu}{d\mu}$. It conveniently reads as

$$d\nu = \frac{d\nu}{d\mu} d\mu.$$

Example 3.2.5 Let $X = [0, 1]$, $\mathcal{M} = \mathcal{B}_{[0,1]}$. Let m be the Lebesgue measure and μ the counting measure on $\mathcal{M}$ defined for all $E \in \mathcal{M}$ as $\mu(E) = \sum_{x \in E} 1$. It is clear that if $\mu(E) = 0$ then $m(E) = \int_E f\, d\mu = 0$ for any f, i.e. $m \ll \mu$. However, there is no such f that $dm \neq f\, d\nu$ since if $E \neq \emptyset$, we have $\mu(E) = \infty$ but $m(E) = b - a$. It follows that μ has no Lebesgue decomposition with respect to m.

3.3 Functions of Bounded Variation

The results of this section can be found in many books on measure theory or in the related chapters of advanced books on analysis. In this case they are particularly based on [36] and [60].

Definition 3.3.1 (Total Variation) Let $F : [a, b] \to \mathbb{R}$ be a function on an interval $[a, b]$.

(i) A *partition* of $[a,b]$ is given by

$$P := \{a = t_0 < t_1 < \ldots < t_n = b\};$$

(ii) The *variation* of F over $[a,b]$ is defined as

$$V(P,F) := \sum_{k=1}^{n} |F(t_k) - F(t_{k-1})|;$$

(iii) The *total variation* of F on $[a,b]$ is defined as

$$TV_{[a,b]}(F) := \sup\{V(P,F) : P \text{ is a partition of } [a,b]\}.$$

Observe that adding more subdivision points to the partition P can only increase but not diminish the value of $V(P,F)$. The total variation of F may assume the value $\{\infty\}$.

Definition 3.3.2 (Functions of Bounded Variation) Let $F : [a,b] \to \mathbb{R}$ be a given function. The function F is said to be of *bounded variation (BV) on* $[a,b]$ if $TV_{[a,b]}(F) < \infty$.

Example 3.3.3 The function $F(x) = x^2 \sin(x^{-1})$ if $x \neq 0$, $F(0) = 0$ is of bounded variation on $[a,b]$ (Figure 3.1a) while the functions

$$F(x) = \begin{cases} \sin(\frac{1}{x}), & x \neq 0, \\ 0, & x = 0 \end{cases} \quad \text{and} \quad F(x) = \begin{cases} x\sin(\frac{1}{x}), & x \neq 0, \\ 0, & x = 0 \end{cases}$$

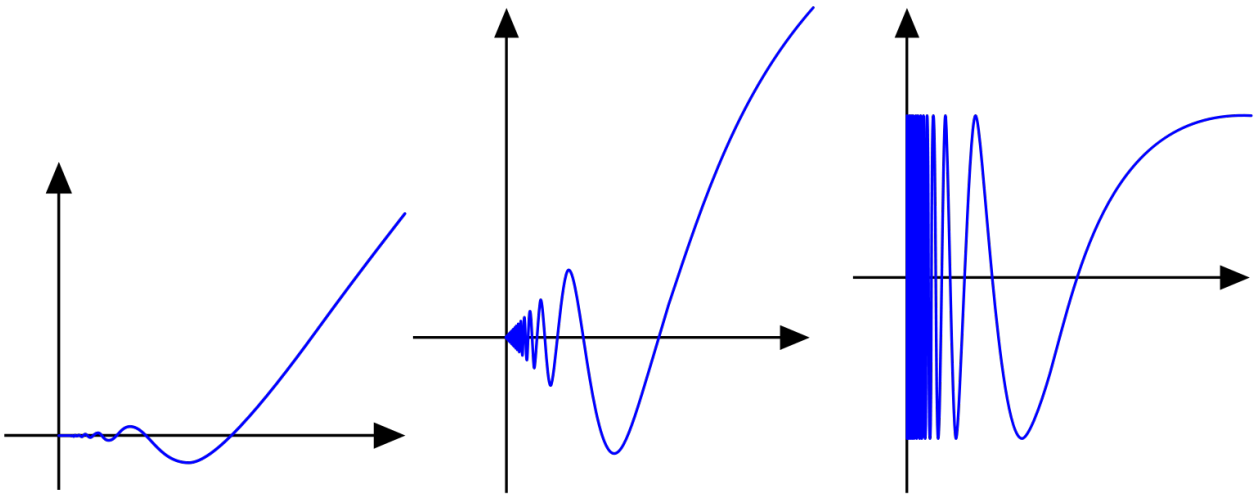
are not in $BV([a,b])$ for $a \leq 0 < b$ or $a < 0 \leq b$ (Figures 3.1b, 3.1c) The factor 1 or x is not enough to damp the oscillation of $\sin(x^{-1})$ as $x \downarrow 0$ whereas x^2 damps sufficiently as Figure 3.1a illustrates.

We state some properties of functions of bounded variation that we will find useful later.

Proposition 3.3.4 (cmp. [36] Lemma 3.26, Thm. 3.27 (b),(d)) Let F be a function of bounded variation defined on $\mathbb{R}$. The following properties hold.

(i) both $TV_{[a,b]}(F) + F$ and $TV_{[a,b]}(F) - F$ are increasing functions;

(ii) F can be written as the difference of some bounded increasing functions F_1, F_2. This implication is also valid in the opposite direction: If F_1, F_2 are bounded and increasing, $F_1 - F_2 = F$, then $F \in BV$.



(a) $x^2 \sin(x^{-1})$ (b) $x \sin(x^{-1})$ (c) $\sin(x^{-1})$

Figure 3.1: An example (and counterexamples) of functions of bounded variation.

Lemma 3.3.5 ([60] VIII §1) Let $F : [a, b] \to \mathbb{R}$ be an increasing function and a limited number of arbitrary points $t_1, t_2, \ldots, t_n \in (a, b)$. Then

$$\begin{aligned}[F(a+) - F(a)] + [F(b) - F(b-)] \\ + \sum_{t=1}^{n} [F(t_k+) - F(t_k-)] \leq F(b) - F(a)\end{aligned} \tag{3.1}$$

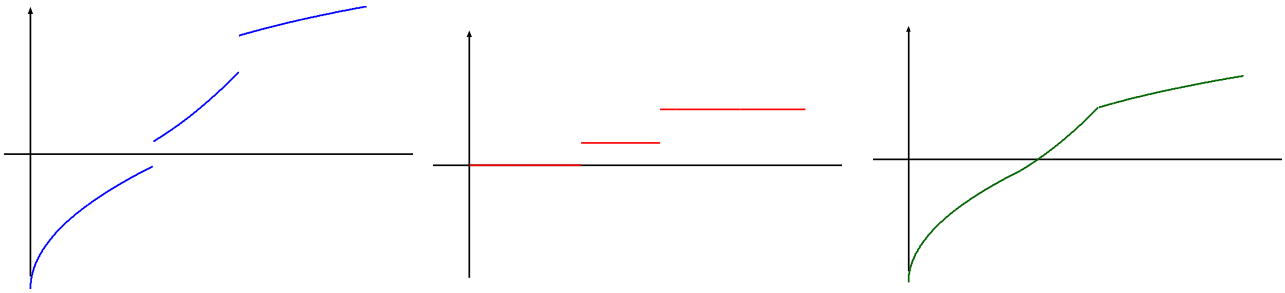
Theorem 3.3.6 ([60] VIII §1, Thm. 1)
Let $F : [a, b] \to \mathbb{R}$ be an increasing function and $t_1, t_2, \ldots \in (a, b)$ the points of discontinuity of F. Then the number of such points is at most countable and

$$\begin{aligned}[F(a+) - F(a)] + [F(b) - F(b-)] \\ + \sum_{t=1}^{\infty} [F(t_k+) - F(t_k-)] \leq F(b) - F(a)\end{aligned} \tag{3.2}$$

The following propositon is an immediate consequence of the preceding theorem and Proposition 3.3.4 (ii).

Proposition 3.3.7 Let $F \in BV([a, b])$. Then the set of points at which F is discontinuous is countable.

Now we look at the left hand side of (3.1) and (3.2) from a different perspective. We look at the "cumulative jumps" of F encountered in a subinterval (a, x).



(a) An increasing F (b) Jump function s (c) $\psi = F - s$

Figure 3.2: An increasing functions with jumps, the corresponding jump function and their difference.

Definition 3.3.8 Let $F : [a, b] \to \mathbb{R}$ be an increasing function. Define the *jump function* $s : [a, b] \to [0, \infty)$ of F by setting

$$s(a) = 0,$$
$$s(x) = [F(a+) - F(a)] + \sum_{t_k < x} [F(t_k+) - F(t_k-)] + [F(x) - F(x-)], \quad a < x \leq b.$$

Obviously, s is an increasing function.

The jump function arising from a discontinuous increasing function is, in fact, a step function, as illustrated in Fig. 3.2a, 3.2b. As the following theorem shows, their difference is a continuous increasing function, depicted in Fig. 3.2c.

Theorem 3.3.9 ([60] VIII §1, Thm. 2) The difference

$$\psi := F - s$$

between an increasing function F and its jump function s is an increasing and continuous function.

PROOF: Let $a \leq x < y \leq b$. The inequality (3.2) can be applied to $[x, y]$ instead of $[a, b]$ and then produces

$$s(y) - s(x) < F(y) - F(x). \tag{3.3}$$

This means that $\psi(x) \leq \psi(y)$, i.e., ψ is an increasing function. Letting $y \downarrow x$ in (3.3) we obtain

$$s(x+) - s(x) \leq F(x+) - F(x) \tag{3.4}$$

From the definition of s it follows that

$$F(x+) - F(x) \leq s(y) - s(x), \quad \forall x < y.$$

After taking the limit $y \downarrow x$ it reads

$$F(x+) - F(x) \leq s(y) - s(x).$$

The last inequality together with (3.4) yields

$$F(x+) - F(x) = s(y) - s(x),$$

that is, $\psi(x+) = \psi(x)$. A similar routine for the left limit shows $\psi(x-) = \psi(x)$, i.e., φ is continuous.■

Proposition 3.3.10 Any function $F \in BV([a,b])$ can be decomposed as

$$F = \psi + s,$$

where ψ is a continuous function of bounded variation and s a jump function defined over the same interval.

PROOF: We know that F can be written as $F_1 - F_2$ with some increasing functions F_1, F_2. Let $x_1, x_2, \ldots$, all in (a,b) be the points of discontinuity of either F_1 or F_2. Consider, for $x \in (a,b]$, the two jump functions s_i, $i = 1, 2$,

$$s_i(x) = [F_i(a+) - F_i(a)] + \sum_{x_k < x}^{\infty} [F_i(x_k+) - F_i(x_k-)] + [F_i(x) - F_i(x-)]$$

with $s_1(a) = s_2(a) = 0$. Let $s(x) = s_1(x) - s_2(x)$. Clearly,

$$s(x) = [F(a+) - F(a)] + \sum_{x_k < x}^{\infty} [F(x_k+) - F(x_k-)] + [F(x) - F(x-)].$$

Note that if we remove from $\{x_k\}$ any points at which F is continuous (it can be shown, in fact, there are no such points) then $s(x)$ remains unchanged. According to Theorem 3.3.9 we know that

the functions $F_1 - s_1$ and $F_2 - s_2$ are continuous and increasing. It follows that $\psi : [a, b] \to \mathbb{R}$,

$$\psi := F - s = (F_1 - s_1) - (F_2 - s_2)$$

is a continuous function of bounded variation. ∎

Definition 3.3.11 A function $r \in BV([a, b])$ is called *singular* if r is continuous and $r' = 0$ almost everywhere in $[a, b]$.

Definition 3.3.12 A function $F : [a, b] \to \mathbb{R}$ is *absolutely continuous (AC)* if for every $\varepsilon > 0$ there exists a $\delta > 0$ so that for any finite set of disjoint intervals $\{(a_j, b_j)\}_{j=1,\dots,N}$ the following holds:

$$\sum_{k=1}^{N} (b_k - a_k) < \delta \quad \Longrightarrow \quad \sum_{k=1}^{N} |F(b_k) - F(a_k)| < \varepsilon.$$

Remark 3.3.13 If a singular function is absolutely continuous, then it must be constant. Conversely, a non-constant singular function is necessarily not absolutely continuous. An example is the Cantor function (See e.g. [36]).

Remark 3.3.14 If F is absolutely continuous on $[a, b]$, then $F \in BV([a, b])$.

Theorem 3.3.15 ([60] IX §6, Thm. 1) Let $F \in BV([a, b])$. If F is continuous, then it can be written as

$$F = \varphi + r,$$

where φ is $AC([a, b])$ and r is a singular function or zero. This representation is unique.

Proof: The derivative F' exists almost everywhere in $[a, b]$ and it is bounded. Set

$$\varphi(x) := F(a) + \int_a^x F'(t)\, dt, \qquad r(x) := F(x) - \varphi(x),$$

where φ is by construction an absolutely continuous function with $\varphi(a) = F(a)$. From the second equation we see that r is continuous, of bounded variation with $r' = F' - F' = 0$ almost everywhere in $[a, b]$. Hence, r is a singular function.

Note that $r \equiv 0$ in $[a, b]$ if and only if F is itself absolutely continuous. It remains to show the uniqueness of the representation. Suppose there are two such representations, we have

$$F(x) = \varphi(x) + r(x) = \varphi_1(x) + r_1(x)$$

for all $x \in [a, b]$, and, equivalently,

$$\varphi(x) - \varphi_1(x) = r(x) - r_1(x).$$

$\varphi - \varphi_1$ is absolutely continuous and, since $\varphi' - \varphi_1' = 0$ almost everwhere, $\varphi - \varphi_1$ is constant. Additionally, $\varphi(a) = \varphi_1(a) = F(a)$. This implies $\varphi(x) \equiv \varphi_1(x)$, and, therefore, $r(x) \equiv r_1(x)$. ∎

Proposition 3.3.16 ([60]) Any function $F \in BV([a, b])$ can be represented as the sum

$$F = \varphi + r + s,$$

where φ is an absolutely continuous function, r is a singular function and s is a jump function.

PROOF: Due to Proposition 3.3.10 we know that F can be decomposed as $F = \psi + s$, where ψ is a continuous function of bounded variation and s is a jump function. According to Theorem 3.3.15, $\psi = \varphi + r$. The sum of the two equations completes the proof. ∎

Definition 3.3.17 (Space of Normalized Functions of Bounded Variations) The *space of normalized functions of bounded variation on* $[a, b]$ is defined by

$$NBV([a, b]) := \{F \in BV \ : \ F \text{ is right continuous and } F(a) = 0\}.$$

3.4 Borel Measures and Normalized Functions of Bounded Variations

Borel Measures

Definition 3.4.1 Let $\mathcal{O}$ be the sets of all open intervals in $\mathbb{R}$. The σ-algebra generated by $\mathcal{O}$, $\sigma(\mathcal{O})$, is called the *Borel σ-algebra*, denoted by $\mathcal{B}_{\mathbb{R}}$. Any element $B \in \mathcal{B}_{\mathbb{R}}$ is called a *Borel set*.

It can be shown that the σ-algebra $\mathcal{B}_{\mathbb{R}}$ is also generated by each of the following sets: the closed intervals, the half-open intervals, the open rays and the closed rays in $\mathbb{R}$.

Definition 3.4.2 A measure defined on $\mathcal{B}_{\mathbb{R}}$ is called a *Borel measure*.

Correspondence between Borel Measures and Functions of NBV

Next we establish a correspondence between Borel measures and functions of NBV. Most results are due to Folland [36] while the formulation of the Riesz Representation Theorem is taken from Luenberger [54].

Theorem 3.4.3 (cmp. [36] 1.16)

(i) If $F : \mathbb{R} \to \mathbb{R}$ is an increasing and right continuous function, there exists a unique measure μ_F defined in $\mathcal{B}_\mathbb{R}$ such that

$$\mu_F((a, b]) = F(b) - F(a) \quad \forall\, a, b \in \mathbb{R}.$$

If G is another such function, then $\mu_F = \mu_G$ if $F(x) - G(x)$ is constant.

(ii) Converesely, if μ is a Borel measure on $\mathbb{R}$ that is finite on all bounded Borel sets, define

$$F(x) := \begin{cases} \mu((0, x]) - F(0), & \text{if } x > 0, \\ F(0), & \text{if } x = 0, \\ -\mu((0, x]) + F(0), & \text{if } x < 0, \end{cases}$$

where $F(0)$ is set arbitrarily. Then the resulting F is increasing and right continuous, and for a given $(a, b]$,

$$\mu((a, b]) = \mu_F((a, b]) = F(b) - F(a).$$

Remark 3.4.4 If $F(x) = x$ in the above theorem, then the associated measure μ_F is called the *Lebesgue measure*, denoted by m and the measure of an interval is simply its length,

$$\mu_F((a, b]) = m((a, b]) = b - a.$$

The domain of m is called the class of *Lebesgue measurable* sets, denoted by $\mathcal{L}$. We will also speak of Lebesgue measure refering to $m|_{\mathcal{B}_\mathbb{R}}$.

Definition 3.4.5 A Borel measure μ on $\mathbb{R}$ is called *regular* if $\mu(K) < \infty$ for every compact K.

It can be shown that a regular Borel measure μ has among all the following properties:

(i) μ is σ-*finite*, i.e. if $\mathbb{R}$ can be written as $\mathbb{R} = \bigcup_{i=1}^{\infty} E_i$, where $E_i \in \mathcal{B}_\mathbb{R}$ and $\mu(E_i) < \infty$ for all $i \in \mathbb{N}$;

(ii) $\mu(E) = \inf\{\mu(U) : U \text{ open}, E \subset U\}$ for every $E \in \mathcal{B}_{\mathbb{R}}$.

Theorem 3.4.6 (cmp. [36] Thm. 3.29) If there exits a regular Borel measure μ on $\mathbb{R}$ and a function F such that $supp\{\mu\} \subset [a,b]$, where

$$supp\{\mu\} := \left\{ \bigcap_{i=1}^{\infty} A_i : A_i \text{ is a closed subset of } [a,b] \text{ and for all } B \subset ([a,b] \setminus A_i)\colon \mu(B) = 0 \right\},$$

and such that $F(x) = \mu((a,x])$, for all $x \in (a,b]$, then $F \in NBV([a,b])$.
Conversely, if $F \in NBV([a,b])$, there is a unique $\mathcal{B}_{[a,b]}$ measure μ_F such that

$$F(x) = \mu_F((a,x]), \quad x \in (a,b].$$

Moreover, $|\mu_F| = \mu_{TV_{[a,b]}(F)}$.

PROOF: (i) Assume that $F(x) = \mu((a,x])$, then $\lim_{x\to\infty} F(x) = \mu((a,\infty)) < \infty$. F is increasing, Proposition 3.3.4 (ii) yields that $F \in BV$. F is right continuous, since for $x_n := y + (1/n)$

$$\mu(a,y]) = \lim_{x \downarrow y} \mu((a,x]) = \mu\left(\bigcap_{i=1}^{\infty} (a, x_n] \right),$$

and also satisfying

$$F(a) = \mu(\{a\}) = 0$$

meaning $F \in NBV([a,b])$.

(ii) If $F \in NBV([a,b])$, due to Proposition 3.3.4 (ii) it can be written $F = F_1 - F_2$ with F_1, F_2 increasing and bounded. From $F_1(a) - F_2(a) = 0$ it follows $F_1(a) = F_2(a) = 0$ and thus both $F_1, F_2 \in NBV([a,b])$. Due to Theorem 3.4.3 F_1, F_2 can be associated with unique measures μ^+, μ^-, respectively. We obtain

$$F(x) = \mu^+((a,x]) - \mu^-((a,x]) =: \mu_F((a,x]) \quad \forall\, x \in (a,b].$$

The equality of $|\mu_F|$ and $\mu_{TV_{[a,b]}(F)}$ states that no matter which particular function F induces the measure, its (absolute) value equals that of the measure induced by the total variation of F. The technical proof is omitted here. ∎

We will make use of the Fundamental Theorem of Calculus which we cite here:

Theorem 3.4.7 Assume a function $F : [a, b] \to \mathbb{R}$. The following are equivalent:

(i) F is absolutely continuous on $[a, b]$;

(ii) $F(x) - F(a) = \int_a^x f(t)\,dt$ for some $f \in L^1([a, b], m)$;

(iii) F is almost everywhere differentiable on $[a, b]$, $F' \in L^1([a, b], m)$ and

$$F(x) - F(a) = \int_a^x F'(t)\,dt.$$

The proof of this important result is omitted here and can be found e.g. in [36].

Theorem 3.4.8 (cmp. [36] Prop 3.30)

(i) If $F \in NBV([a, b])$, then $F' \in L^1([a, b], m)$;

(ii) $\mu_F \perp m$ iff $F' = 0$ *a.e.* in $(a, b]$;

(iii) $\mu_F \ll m$ iff $F(x) = \int_a^x F'(t)\,dt$.

We need one more definition of a special kind of measure to be able to state the central result of this exposition: the decompostion of a regular Borel measure in its absolutely continuous, singular and discrete parts.

Definition 3.4.9 (Discrete measure) Let μ be a regular Borel measure on $[a, b] \subset \mathbb{R}$. It is called discrete if there is a countable set $\{x_j\} \subset [a, b]$ and real numbers $c_j > 0$ such that $\sum c_j < \infty$ and

$$\mu(E) = \sum_j c_j \delta_{x_j},$$

where δ_{x_j} is the Dirac measure at x_j (recall Definition 3.1.6).

Proposition 3.4.10 Let μ be a regular Borel measure on $\mathbb{R}$. Then it can be decomposed as

$$\mu = \mu_d + \mu_{ac} + \mu_{sc},$$

where μ_d is a discrete measure, μ_{ac} is an absolutely continuous measure (i.e., $\mu_{ac} \ll m$) and μ_{sc} is a singular continuous measure (i.e., $\mu_{sc} \perp m$). This decomposition is unique.

PROOF: According to Theorem 3.4.3 we know that μ can be uniquely associated with a function F of normalized bounded variation. Due to Proposition 3.3.16, F can be written as $F_{ac} + F_{sc} + F_d$, where F_{ac} is an absolutely continuous function, F_{sc} is a singular function and F_d is a jump function, and this representation is unique. Each of these functions is itself a function of bounded variation. Again, applying Theorem 3.4.3 we are able to find for each of F_{ac}, F_{sc}, F_d a regular Borel measure $\mu_d, \mu_{ac}, \mu_{sc}$, respectively.

The properties $\mu_{ac} \ll m$ and $\mu_{sc} \perp m$ follow from Theorem 3.4.8 applied to F_{ac} and F_{sc}, respectively.■

Remark 3.4.11 In fact, it can be easily seen that also $\mu_d \perp m$: Define $\hat{E} := \{x_j\}$, the set of discontinuities of F. Then $\mu(\hat{E}) = \mu_d(\hat{E}) = \sum_j c_j$ whereas $m(\hat{E}) = 0$. Choose, on the contrary a set of disjoint intervals $\tilde{E}_i$ such that $\cup\tilde{E}_i = [a,b] \setminus \hat{E}$. Then $m(\cup\tilde{E}_i) = b - a$ whereas $\mu_d(\cup\tilde{E}_i) = 0$.

Theorem 3.4.12 (Riesz Representation Theorem, [54] §5.5 Thm. 1)
Let $F : C([a,b];\mathbb{R}) \to \mathbb{R}$ be a bounded linear functional. Then there exists a unique function $v \in NBV([a,b])$, such that for all $y \in C([a,b];\mathbb{R})$

$$F(y) = \int_{(a,b]} y\, dv$$

and the norm of F is the total variation of v, $TV_{[a,b]}(v)$. Conversely, every function $v \in NBV([a,b])$ defines uniquely a bounded linear functional F on $C([a,b];\mathbb{R})$.

Remark 3.4.13 The proof which is omitted here essentially hinges on the extension of F from $C([a,b];\mathbb{R})$ to the space of bounded linear functions $[a,b]$ as a consequence of the Hahn-Banach theorem (e.g. [54] §5.4 Thm. 1).

It should be noted that the bounded linear functional F could be represented by a function $v \in BV([a,b])$ in the formulation of the above theorem. However, this representation would not be necessarily unique. Thus, the theorem was formulated for functions v of $NBV([a,b])$ as a subspace of $BV([a,b])$ which provides for uniqueness of v and, at the same time, the unique association between $NBV([a,b])$ and the dual of $C([a,b];\mathbb{R})$.

In the view of Theorem 3.4.3, this allows us to consider the dual of $C([a,b];\mathbb{R})$, denoted by $C^*([a,b];\mathbb{R})$, as the space of Borel signed measures.

Chapter 4

Introducing the SEIR Problem

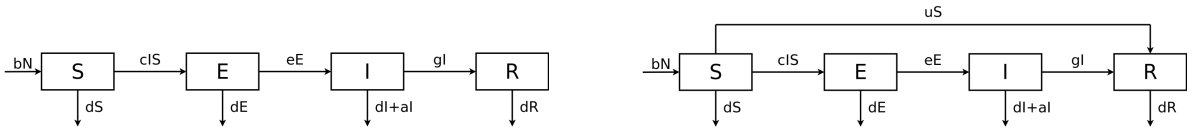
Optimal control can be used to determine vaccination policies for various infectious diseases. Many times the so called compartmental models are used to describe the epidemic in question. In this thesis the problem is presented as a rich playground for verifying the numerical solutions against the analytically derived ones.

We will consider state constraints of various kind. In Chapter 5, the problem is coupled with a pure state constraint and control-quadratic L^2 cost functional. In Chapter 7, the problem is treated with a cost functional of L^1 type, linear with respect to control variable, and a mixed state constraint.

4.1 Model Description

The SEIR model is a *compartmental model* dividing the total population N into four different compartments regarding the epidemic. Those compartments are *susceptible* (S), *exposed* (E), *infectious* (I), and *recovered* (or immunized by vaccination) (R).

First, we introduce the "natural" (uncontrolled) dynamics of the SEIR model. Those in the S compartment are susceptible to contracting the disease. A person who is infected but is currently in latency is in the E compartment. Infectious individuals are in the I compartment and immune ones are in the R compartment. Any newborn is considered susceptible. A susceptible individual becomes exposed when in contact with infectious individuals by what is known as horizontal transmission (via direct or indirect contact with infected individuals, see [61] for a more complete description). The exposed ones may die of natural causes or become infectious. The infectious ones can either die or recover completely. Finally, all individuals who recovered (those susceptible who were vaccinated or those who recovered from the disease) are considered immune. These dependencies are depicted in



(a) without vaccination (b) with vaccination parameter u

Figure 4.1: The SEIR model with and without vaccination

Figure 4.1 (a). The total population at any time t is $N(t) = S(t) + E(t) + I(t) + R(t)$.

The control $u(t)$ is the rate of vaccination taking on values in $[0, 1]$. Only susceptible individuals are vaccinated (we have $u(t) = 1$ if at a given instant all susceptible individuals are vaccinated.) We assume that every vaccinated individual becomes immune, that is, an individual in the compartment S, treated with vaccine, proceeds to the R compartment. The effect of vaccination is shown in Figure 4.1 (b).

We shall look at the evolution of the disease over a certain period of time T. The parameters describing the population and the disease transmission, assumed constant over the period of time of interest, are the following. The birth rate of the population is b while d denotes the natural death rate. The rate at which the exposed individuals become infectious is e, g is the rate at which infectious individuals recover and a denotes the death rate due to the disease. The rate of transmission is described by the number of contacts between susceptible and infectious individuals. If c is the incidence coefficient of horizontal transmission, such rate is $cS(t)I(t)$ (see Table 4.1 below).

Taking all the above considerations into account we are led to the following dynamical system:

$$\dot{S}(t) = bN(t) - dS(t) - cS(t)I(t) - u(t)S(t) \tag{4.1}$$
$$\dot{E}(t) = cS(t)I(t) - (e + d)E(t) \tag{4.2}$$
$$\dot{I}(t) = eE(t) - (g + a + d)I(t) \tag{4.3}$$
$$\dot{R}(t) = gI(t) - dR(t) + u(t)S(t) \tag{4.4}$$
$$\dot{N}(t) = (b - d)N(t) - aI(t) \tag{4.5}$$

with the initial conditions $S(0) = S_0$, $E(0) = E_0$, $I(0) = I_0$, $R(0) = R_0$ and $N(0) = N_0$.

The differential equation for the recovered compartment (R) can be removed since the state variable R only appears in the corresponding differential equation and the number of recovered individual at each instant t is obtained from $R(t) = N(t) - S(t) + E(t) + I(t)$. However, in other chapters we will be interested to count the number of vaccinated individuals, therefore we introduce an extra variable W and the differential equation $\dot{W}(t) = u(t)S(t)$ with the initial condition $W(0) = 0$. As far as the optimal control problem is concerned, this new differential equation is redundant.

4.2 L^2 vs. L^1 Cost Functional

There is more than one candidate for the choice of a cost functional for the control problem. In Chapter 5 the same cost quadratic functional is used as it was introduced by Neilan and Lenhart in [61] (and also in [9]):

$$J_2(x,u) = \int_0^T \left(AI(t) + u^2(t)\right)\, dt.$$

In contrast, the cost functional appearing in Chapter 7 is of L^1 type.

$$J_1(x,u) = \int_0^T \left(A_1 I(t) + B_1 u(t)\right)\, dt.$$

Note the difference between the constants A and A_1, B_1 in the L^2 and L^1 case. The convexity of $J_2(x,u)$ with respect to u is advantageous for the numerical approach, since it allows to express the control variable in terms of the state and the adjoint variable. In both cases, the cost functional is a weighted sum of the overall cost of caring for the infected individuals and the cost of vaccination. Observe, however, that in the case of $J_2(x,u)$ the cost of vaccination will depend on u^2, a small quantity compared to u which takes values less than 1. In this respect, $J_1(x,u)$ is a more realistic cost functional.

4.3 Numerical Setup

In Table 4.1 we present the values of the parameters and constants used in all our simulations. Such values are exactly as in [61]. The values of S_0, E_0, I_0, N_0 and W_0 appear in the last lines of the table.

For our simulations we use the Imperial College London Optimal Control Software – ICLOCS – version 0.1*b* [34]. ICLOCS is an optimal control interface, implemented in Matlab, for solving optimal control problems. It calls IPOPT (Interior Point OPTimizer), an open-source software package for large-scale nonlinear optimization [76]. See also [64] for a more detailed explanation.

Considering a time interval of 20 years ($T = 20$), a time-grid with 10000 nodes was created, i.e., for $t \in [0, 20]$ we get $\Delta t = 0.002$. Since our problem is solved by the direct method we impose an acceptable convergence tolerance at each step of $\varepsilon_{rel} = 10^{-9}$. In this respect one may consult [9].

Table 4.1: Parameters with their clinically approved values and constants as in [61].

Parameter	Description	Value
b	natural birth rate	0.525
d	natural death rate	0.5
c	incidence coefficient	0.001
e	exposed to infectious rate	0.5
g	recovery rate	0.1
a	disease induced death rate	0.2
T	number of years	20
S_0	initial susceptible population	1000
E_0	initial exposed population	100
I_0	initial infected population	50
R_0	initial recovered population	15
N_0	initial population	1165
W_0	initial vaccinated population	0

Part II

New Contributions

Chapter 5

The SEIR Problem with State Contraints and L^2 Cost

In this chapter we add state constraints to the L^2 case of the SEIR model described in Chapter 4. A first thought would be to keep a pointwise upper bound on the number of infectious individuals. However, this would be a state constraint of *order higher than one* known to be hard to treat numerically (and theoretically). We choose to impose an upper bound on S which is a *first order* state constraint (as we will show). It turns out that our choice is of relevance in practical terms. Since the spreading of the disease is given by $cS(t)I(t)$, it is reasonable to expect that the number of infectious individuals will be driven down because of the upper bound on the number of susceptible individuals. Note that the number of susceptible individuals will certainly increase given that any newborn is considered susceptible but, after vaccination a susceptible individual becomes immune. The translation of the upper bound on the number of susceptible individuals into mathematical terms is the state constraint $S(t) \leq S_{max}$.

This work was published in [48].

5.1 Introduction

We focus on a general state constrained optimal control problem with dynamics linear with respect to control. Our problem of interest is the following fixed-time problem:

$$
(\mathrm{P}) \quad \begin{cases} \text{Minimize } l(x(T)) + \displaystyle\int_0^T L(x(t), u(t))\, dt \\ \text{subject to} \\ \quad \dot{x}(t) \quad = f_1(x(t)) + g(x(t))u(t) \quad \text{for a.e. } t \in [0,T], \\ \quad h(x(t)) \quad \leq 0 \quad \text{for all } t \in [0,T], \\ \quad u(t) \quad \in U \quad \text{for a.e. } t \in [0,T], \\ \quad x(0) \quad = x_0, \\ \quad x(T) \quad \in \mathbb{R}^n. \end{cases}
$$

Here the state x takes values in $\mathbb{R}^n$ while the control $u \in \mathbb{R}^k$ and U, the control set, is a subset of $\mathbb{R}^k$. As for the functions we have $l : \mathbb{R}^n \to \mathbb{R}$, $L : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}$, $f_1 : \mathbb{R}^n \to \mathbb{R}^n$, $g : \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^k$ and $h : \mathbb{R}^n \to \mathbb{R}$.

The next section states auxiliary results concerning (P), including necessary conditions.

5.2 Necessary Conditions

Take any admissible process (x, u) for (P). Set $h^0(t, x, u) = h(x(t))$ and $h^1(x, u) = \dfrac{dh}{dt}(x(t))$. With respect to the dynamics we have

$$
h^1(x, u) = \nabla_x h^0(x)\dot{x} = \left\langle \frac{\partial h}{\partial x}(x), f_1(x) + g(x)u \right\rangle.
$$

If for all $t \in [0, T]$ we have $\dfrac{\partial h^1}{\partial u}(x, u) = \left\langle \dfrac{\partial h}{\partial x}(x), g(x) \right\rangle \neq 0$ then we say that the state constraint is of *order one*. (For a general definition of the order of a constraint see [42]).

Let (x^*, u^*) be a reference process for (P) and ϵ a given parameter. We impose the following condition on the data of (P).

(H1) The function $u \to L(x, u)$ is continuous on U for all $x \in \mathbb{R}^n$;

(H2) The functions $x \to f_1(x)$, $x \to g(x)$, $x \to h(x)$ and $x \to L(x, u)$ are continuously differentiable

on $x^*(t) + \epsilon B$ for all $u \in U$;

(H3) The function l is Lipschitz continuous on $x^*(T) + \epsilon B$;

(H4) The set U is compact.

Regarding the existence of solution for (P), we refer to Theorem 23.11 in [18] which asserts that (P) has a solution if (H1)–(H4) are satisfied and an admissible solution exists.

Suppose that (x^*, u^*) is a local strong minimum. Under our assumptions, Theorem 9.3.1 in [75] applies asserting the existence of an absolutely continuous function p, a scalar λ and a measure $\mu \in C^{\oplus}([0,T])$ such that

(i)
$$(p, \lambda, \mu) \neq (0, 0, 0), \tag{5.1}$$

(ii)
$$-\dot{p}(t) = f_{1,x}(x^*(t))^T q(t) + u^*(t) g_x(x^*(t))^T q(t) - \lambda L_x(x^*(t), u^*(t)), \tag{5.2}$$

(iii)
$$\langle g(x^*(t))u^*(t), q(t)\rangle - \lambda L(x^*(t), u^*(t)) \geq \langle g(x^*(t))u, q(t)\rangle - \lambda L(x^*(t), u) \quad \forall u \in U, \tag{5.3}$$

(iv)
$$-q(T) = 0, \tag{5.4}$$

(v)
$$supp\{\mu\} \subset \{t : \ h(x^*(t)) = 0\}, \tag{5.5}$$

where
$$q(t) = p(t) + \int_{[0,t)} \nabla h\left(x^*(s)\right) \, \mu(ds), \quad q(T) = p(T) + \int_{[0,T]} \nabla h\left(x^*(s)\right) \, \mu(ds).$$

The function q is a bounded variation function.

5.3 The SEIR Problem

Our problem of interest is now (P_S):

$$(P_S)\quad \begin{cases} \text{Minimize } \displaystyle\int_0^T \left(AI(t)+u^2(t)\right)\,dt \\ \text{subject to} \\ \quad \dot{S}(t) &= bN(t)-dS(t)-cS(t)I(t)-u(t)S(t), \\ \quad \dot{E}(t) &= cS(t)I(t)-(e+d)E(t), \\ \quad \dot{I}(t) &= eE(t)-(g+a+d)I(t), \\ \quad \dot{N}(t) &= (b-d)N(t)-aI(t), \\ \quad \dot{W}(t) &= u(t)S(t), \\ \quad S(t) &\le S_{max}, \\ \quad u(t) &\in [0,1] \quad \text{for a.e. } t\in[0,T], \\ \quad S(0) &= S_0,\ E(0)=E_0,\ I(0)=I_0,\ N(0)=N_0,\ W(0)=W_0, \end{cases}$$

whereas the function W, not actually part of the control problem, counts the number of vaccinated individuals. The total number of vaccinated persons is then $W(T)$.

This problem is in the form of (P) as it can be seen by setting

$$x(t)=(S(t),E(t),I(t),N(t)),\quad \tilde{A}=(0,0,A,0),\quad C=(1,0,0,0),$$

$$A_1=\begin{bmatrix} -d & 0 & 0 & b \\ 0 & -(e+d) & 0 & 0 \\ 0 & e & -(g+a+d) & 0 \\ 0 & 0 & -a & b-d \end{bmatrix},\quad B=\begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

and defining $l(x)=0$, $L(x,u)=\langle \tilde{A},x\rangle+u^2$, $f_1(x)=A_1x+c(-SI,SI,0,0)^T$, $g(x)=Bx$ and $h(x)=\langle C,x\rangle - S_{max}=S-S_{max}$ for some fixed $S_{max}>S(0)$.

Note that (P_S) has free end states, a quadratic cost with respect to u and that the differential equation $\dot{x}=f_1(x)+g(x)u$ is affine in the control and nonlinear in the state x due to the term f_1.

The initial values we work with are given in Table 4.1 in Chapter 4. In the view of the optimal trajectory of P_S associated with these initial values (Figure 5.1) it is simple to determine some upper and lower bounds which the trajectory never exceedes (nor reaches); i.e. there are positive constants $U_S, L_S, U_N, L_N, U_E, U_I$ such that for all $t\in[0,T]$

$$(S(t),E(t),I(t),N(t))\in(L_S,U_S)\times(0,U_E)\times(0,U_I)\times(L_N,U_N)\,.$$

For the discussion of normality in the following Section 5.4 it is sufficient to draw these bounds at $L_S = 500$, $U_S = 1900$, $U_E = E(0) = 100$, $U_I = 55$, $L_N = N(0) = 1165$, $U_N = 1900$.

5.4 Normality

There exist various approaches in the literature for proving normality [1] which go back, for example, to H. Maurer, H. Frankowska, R. Vinter and others. We will make use of the following *inward pointing condition* from the work of F. Rampazzo and R. Vinter [68]: Let $\xi = (S, E, I, N)$ and make the assumption

(RV-H) There exist constants $\varepsilon > 0, \gamma, \delta$ and a continuous function $\nu : [a, b] \times \mathbb{R}^n \to [0, 1]$ such that, if $(t, \xi) \in [a, b] \times \mathbb{R}^n$, $|\xi - x^*(t)| \leq \varepsilon$ and $h(x^*(t)) > -\delta$, and the condition

$$\nabla_x h(\xi) \cdot [f_1(\xi) + g(\xi)\nu(t, \xi)] < -\gamma$$

is satisfied.

We verify **(RV-H)** with $\xi := (S, E, I, N)$ (the components of this vector to be defined shortly), $\nu(t, \xi) \equiv 1$, $\delta = 2$, $\nabla h = (1, 0, 0, 0)$ (note that ∇h is not dependent on x). Thus,

$$\nabla_x h(\xi) \cdot [f_1(\xi) + g(\xi)\nu(t, \xi)] = \dot{S}(\xi, \nu).$$

We need to show that $\exists\, \gamma > 0$ such that, $\forall t \in [0, T]$, the inequality

$$\dot{S}(\xi, 1) = bN - dS - cSI - S < -\gamma \tag{5.6}$$

is satisfied. In fact, it is sufficient to evaluate (5.6) on the boundary interval $[t_e, T]$ only.

We assume that $cS(t)I(t)$ is neglectably small due to $c = 0.001$, $\min N(t) = 1500$, $\max S(t) = 1100$ for $t \in [12, 20]$ and, further, that the inward pointing condition has to be valid in the δ-tube around S. We can validate **(RV-H)** taking into account numerical simulations (see Section 5.5) and considering the worst case for (5.6), i.e., we choose $\xi = (1098, 3, 2, 1900)$, we obtain

$$0.525 \cdot 1900 - 0.5 \cdot 1098 - 0 - 1098 = -649.50 < 0.$$

Therefore Theorem 4.1 in [68] allows us to conclude that problem P_S is normal..

[1] This means that the necessary condition (i)-(iii) of Section 5.2 can be written with $\lambda = 1$.

5.5 Discussion of Necessary Conditions for (P_S)

Let us apply the necessary conditions from Section 5.2 to (P_S) with $\lambda = 1$. Consider $q = (q_s, q_e, q_i, q_n)$ and analogously $p = (p_s, p_e, p_i, p_n)$ to be the multipliers for $x^*(t) = (S^*(t), E^*(t), I^*(t), N^*(t))$.

If $u^*(t) \in (0,1)$, the Weierstrass Condition (5.3) yields $\langle g(x^*(t))u^*, q(t)\rangle - u^{*2} \geq \langle g(x^*(t))u, q(t)\rangle - u^2$ for all $u \in [0,1]$. Taking into account that $g(x^*(t)) = -(S^*(t), 0, 0, 0)$, we deduce that

$$u^*(t) = -\frac{q_s(t)S^*(t)}{2}. \tag{5.7}$$

Since $u^*(t)$ may be 0 or 1 or in $(0,1)$, we conclude that

$$u^*(t) = \max\left\{0, \min\left\{1, -\frac{q_s(t)S^*(t)}{2}\right\}\right\}. \tag{5.8}$$

Suppose now that $\left[t_0^b, t_1^b\right]$ is a boundary interval, as defined in Section 2.2. Then for t on this interval we have $S^*(t) = S_{max}$ and, consequently,

$$\dot{S}^*(t) = bN^*(t) - dS^*(t) - cS^*(t)I^*(t) - u^*(t)S^*(t) = 0.$$

It follows then that for $t \in \left[t_0^b, t_1^b\right]$ we get

$$u^*(t) = b\frac{N^*(t)}{S^*(t)} - d - cI^*(t). \tag{5.9}$$

Recall now that $q(t) = p(t) + \int_{[0,t)} \nabla h(x^*(t))\, \mu(ds)$, $\nabla h(x^*(t)) = (1,0,0,0)$ and

$$\int_{[0,t)} (1,0,0,0)\, \mu(ds) = \left(\int_{[0,t)} \mu(ds), 0, 0, 0\right).$$

Thus we have

$$\begin{aligned} q_s(t) &= p_s(t) + \int_{[0,t)} \mu(ds), \\ q_e(t) &= p_e(t), \\ q_i(t) &= p_i(t), \\ q_n(t) &= p_n(t). \end{aligned}$$

Next we explore *regularity properties of the multipliers*. Viewing the preconditions for Lemma 5.7 in [73], in the light of (H1)-(H2), it remains to verify that:

(a) U is a closed, convex, time-invariant set,

(b) for every $t \in [0,T]$ and every $u^* \in (0,1)$

$$\begin{aligned} g(x^*(t))^T \nabla h(x^*(t)) &= Bx^*(t)^T(1,0,0,0) \\ &= -S^*(t) \neq 0 = N_U(u^*(t)), \end{aligned}$$

(c) the following strong convexity condition on $u \to L(t,x,u)$ holds on a tube

$$\Omega = \left\{(t,x) \in [0,T] \times \mathbb{R}^4 \ : \ |x - x^*(t)| \leq \varepsilon\right\}$$

for a constant $\sigma > 0$, $u_1, u_2 \in U$ and any $\lambda \in (0,1)$:

$$AI + \left[(1-\lambda)u_1 + \lambda u_2\right]^2 \leq (1-\lambda)\left[AI + u_1^2\right] + \lambda\left[AI + u_2^2\right] - \frac{1}{2}\sigma\lambda(1-\lambda)\left|u_1 - u_2\right|^2.$$

Indeed, after reformulation one has

$$(\lambda^2 - \lambda)\left[u_1^2 + u_2^2\right] - 2(\lambda^2 - \lambda)u_1u_2 \leq -\frac{1}{2}\sigma(\lambda - \lambda^2)\left|u_1 - u_2\right|^2,$$

which, after division by $\lambda^2 - \lambda < 0$, yields $0 \geq -\frac{1}{2}\sigma$, i.e., (c) is satisfied for any $\sigma > 0$.

Thus, Lemmas 5.7 and 6.1 in [73] allow to state the Lipschitz continuity of the integral of the measure

$$\int_{t_1}^{t_2} \mu(d\sigma) \leq K\left|t_1 - t_2\right|$$

for some $K > 0$ and all $[t_1, t_2] \subset [0,T]$. Our numerical results show that the optimal trajectory x^* has only one boundary interval $[t_e, T]$ with the entry point $t_e \in (0,T]$. From the above we conclude that μ is absolutely continuous with respect to the Lebesgue measure in $[t_e, T)$ and may write

$$\int_0^t \nu(\sigma)\, d\sigma = \int_{[t_e,t]} \mu(d\sigma) \qquad \text{for } t \in [t_e, T], \tag{5.10}$$

where ν is an integrable function (see Proposition 3.4.10).

Consequently, q_s is absolutely continuous on $[0,T)$ and $\dot{q}_s(t) = \dot{p}_s(t) + \nu(t)$. It is now a simple matter to see that

$$\dot{q}_s(t) = (d + cI^*(t) + u^*(t))q_s(t) - \nu(t) + cI^*(t)q_e(t).$$

Let us now concentrate on a boundary interval. Taking the above expression of $\dot{q}_s$ we deduce with

the help of (5.8) and (5.9) that

$$\nu(t) = -(d + cI^*(t) + u^*(t))q_s(t) - cI^*(t)q_e(t) + 2c\frac{\dot{I}^*(t)}{S_{max}} - \frac{2b\dot{N}^*(t)}{S_{max}^2}. \tag{5.11}$$

The function ν is indeed defined in the whole interval but it is $\nu(t) = 0$ in any interior interval.

Next we will call ν our *analytical multiplier* considering that it is defined as in (5.11) for all t. Such ν will then be compared with its computed counterpart. Finally, it is worth mentioning that q_s may have a jump when $t = T$ and that $p_e(t) = q_e(t)$, $p_i(t) = q_i(t)$, $p_n(t) = q_n(t)$ and $p_e(T) = p_i(T) = p_n(T) = 0$.

We now show the numerical simulations of ($\mathrm{P_S}$). We consider the state constraint $S(t) \leq S_{max}$ with $S_{max} = 1100$.

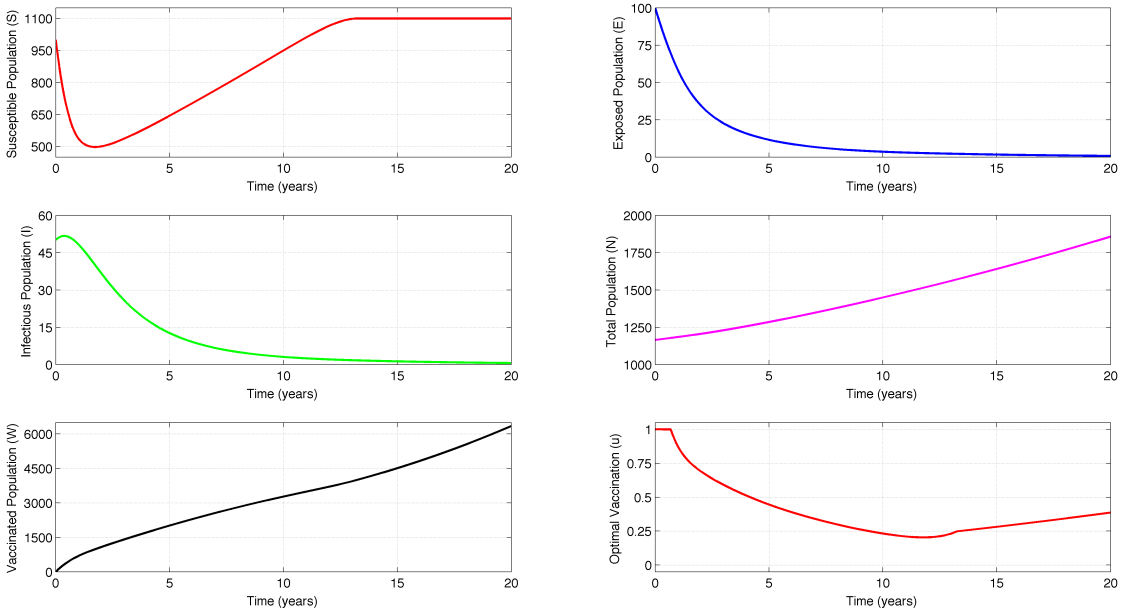


Figure 5.1: The optimal trajectories and optimal vaccination rate for ($\mathrm{P_S}$p[)

Observations:

Notably, the state constraint on S has one boundary interval wich includes $t = T$. (*top-left*)

Control u is 1 in the initial period of time (until I reaches its peak) then becomes singular (*bottom-right vs. center-left*)

In the boundary interval the control remains singular, at the entry point ($t = 13.2$), the control is nondifferentiable (*bottom-right*)

About 6 345 individuals were vaccinated during the whole period. Figure 5.1 shows that the computed optimal control is 1 in the beginning dropping to approximately 0.2 and increasing from then on to

keep the number of susceptible individuals equal or below 1 100. Observe that the state constraint has a boundary interval and that the state constraint is active at the end point (i.e., $S(20) = 1\,100$). The multiplier associated with the S variable, p_s, is not 0 when $T = 20$, as shown in Figure 5.2. This behaviour can be explained since the measure μ has an atom at $t = T$ although it is absolutely continuous with respect to the Lebesgue measure on $[0, T)$.

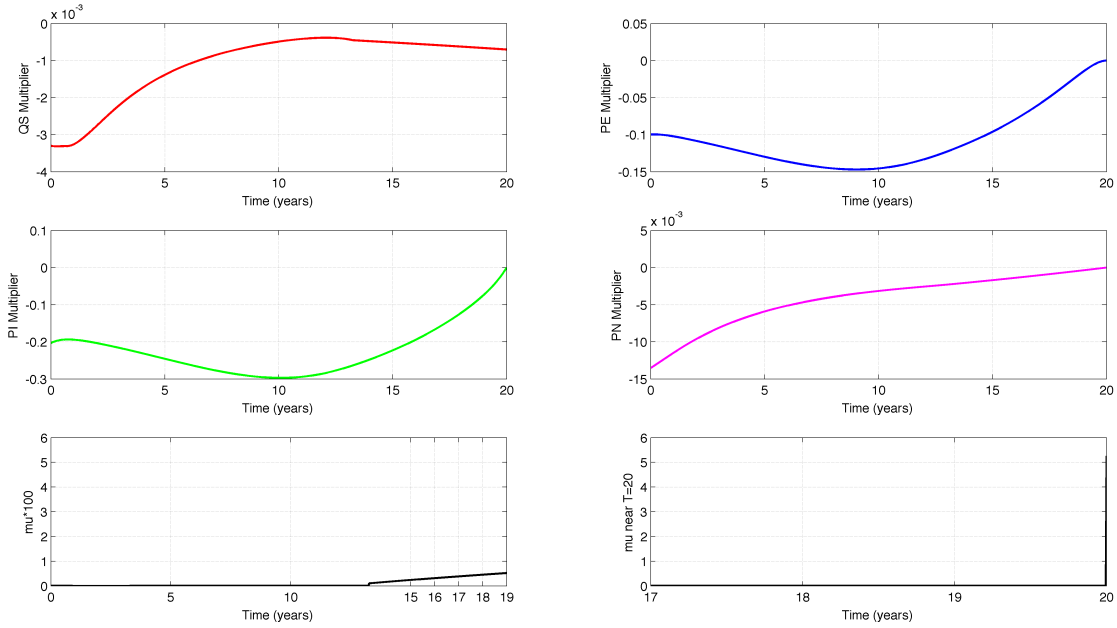


Figure 5.2: The adjoint multipliers for (P_S).

Observations:

The first four subgraphs (from above) show multipliers $q_s = p_s + \eta, p_e, p_i, p_n$ (*clockwise*). Note that $p_e(T) = p_i(T) = p_n(T) = 0$. Also the condition $q_s(T) = 0$ is met, however, not visible at $T = 20$ (compare vs. the singular jump of μ at $T = 20$, *bottom-right*)

Analytical multiplier $\eta(t) = \int_0^t \nu(s)\, ds$ (cf. (5.10) and (5.9)) vs. the scaled computational multiplier (*bottom-left*).

Analytical multiplier $\eta(t)$ vs. the computational multiplier with focus on the terminal interval $[17, 20]$ and the jump present at $T = 20$. (*bottom-right*).

To validate the numerical solution we first use (5.7). As shown in the top left graph of Figure 5.3, the computed optimal control satisfies (5.8). In the top right graph of Figure 5.3, we also show that the computed optimal control matches the control defined by (5.9) when the state constraint is active. We go a step further and compare the multipliers ν computed by ICLOCS with the analytical (5.11). This comparison is shown in the bottom right graph of Figure 5.3. Indeed, we have a match except for $t = T$ where the numerical multiplier ν has a jump at $t = T$, as seen in Figure 5.2. For the sake of completeness we show the graph of the multipliers q_s and p_s in the bottom left graph of 5.3. Recall

that (5.4) asserts $q_s(T) = 0$. It is clear that q_s has a jump at T due to the atom of the measure μ.

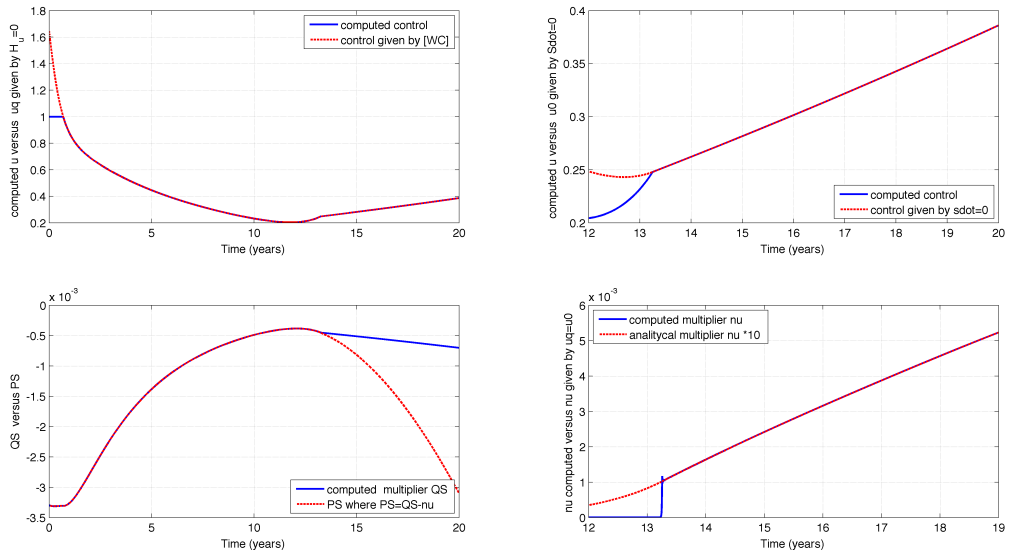


Figure 5.3: The adjoint multipliers for (P_S).

Observations:

Optimal control u^* from (5.8) vs. from Weierstrass Condition (*top-left*).

Optimal control u^* from (5.8) vs. control in the boundary interval (5.9) (*top-right*).

Multipliers q_s and p_s (*bottom-left*).

Analytical multiplier $\eta(t) = \int_0^t \nu(s)\, ds$ (cf. (5.10) and (5.11)) vs. the computational multiplier with focus on the boundary interval $[13.2, 19]$ (*bottom-right*).

5.6 Conclusion

We found this problem very appropriate to test for absolutely continuous state constraint multipliers. Regularity was, indeed, proven for the time period $[0, T)$. However, even with the first order state constraint of this problem we needed to be extremely careful as it turns out that at $t = T$ the measure has an atom. This makes the application of exact penalization difficult. Besides, it would be interesting to find an analytical way of determining the point where the state constraints touches the boundary. Also, stability of the solution with respect to the parameters should be studied. The last question refers to second order conditions sufficient conditions as in, for example, [57], [55] and [63].

Chapter 6

Exact Penalization for State Constrained Problems

Measures as multipliers associated with the state constraint in the Maximum Principle (see for example [75]) are a source of hardship both analytically and numerically. It is thus natural to ask if there exists any class of problems with state constraints where such measures are *well behaved* in the sense that they can be absolutely continuous with respect to the Lebesgue measure. In this chapter we investigate such question. We explore exact penalization techniques to see how the maximum principle would look like and we discuss the difficulties concerning the validation of such result. Notably, we show that a Maximum Principle without a measure would be possible if a certain condition were valid which we call a *hypothetical* condition (HH) (see below.) We determine a sufficient condition for (HH) to hold and test our results using the simple problem of the previous chapter. Taking into account the analysis done in Chapter 5, we know that the measure associated with the state constraint is absolutely continuous inside the interval but has an atom at the end point. Thus our problem, however simple, fails to validate our results.

The results presented in this chapter are published in [47].

6.1 Exact Penalization

Consider the following problem

$$(P) \quad \begin{cases} \text{Minimize } l(x(a), x(b)) \\ \text{subject to} \\ \dot{x}(t) = f(t, x(t), u(t)) \quad \text{a.e. } t \in [a, b] \\ h(x(t)) \leq 0 \quad \text{for all } \ t \in [a, b] \\ u(t) \in U \quad \text{a.e. } t \in [a, b] \\ (x(a), x(b)) \in E. \end{cases}$$

Again, the function $f : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^n$ describes the system dynamics and $h : \mathbb{R}^n \to \mathbb{R}$ is the functional defining the pure state constraint. Furthermore, the set $E \subset \mathbb{R}^n \times \mathbb{R}^n$ and $l : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ specify the endpoint constraints and the cost. The set U defines the set control constraints. Observe that we introduce a simplification by assuming that h is a function of x alone.

As introduced in Chapter 2, this problem involves a measurable control function u and an absolutely continuous function x. Let (x^*, u^*) be a strong local maximum as defined in Chapter 2. We consider the following *basic hypotheses* on the problem data which make reference to (x^*, u^*) and a scalar $\varepsilon > 0$:

(H1) The function $t \longrightarrow f(t, x, u)$ is $\mathcal{L}$-measurable for all x and u.

(H2) There exists a constant $K_l > 0$ such that

$$|l(x_a, x_b) - l(x'_a, x'_b)| \leq K_l |(x_a, x_b) - (x'_a, x'_b)|$$

for all (x_a, x_b), (x'_a, x'_b) such that $x_a, x'_a \in x^*(a) + \varepsilon \bar{B}$, $x_b, x'_b \in x^*(b) + \varepsilon \bar{B}$.

(H3) The set E is closed.

(H4) The function h is continuously differentiable on the tube $\Omega = \left\{x \in \mathbb{R}^n \ : \ x \in x^*(t) + \varepsilon \overline{B}\right\}$ and $\nabla h(x) \neq 0$ for any x such that $h(x) = 0$.

(H5) There exist k_x^f and k_x^f such that, for all $u, u' \in \mathbb{R}^k$ and all $x, x' \in x^*(t) + \varepsilon \bar{B}$, we have

$$|f(t, x, u) - f(t, x', u')| \leq k_x^f |x - x'| + k_u^f |u - u'|$$

for almost every $t \in [a, b]$.

(H6) The set U is compact.

The following lemma will be important in the subsequent development.

Lemma 6.1.1 Let function h satisfy (H4). Define the set $S := \{x \in \mathbb{R}^n : \ h(x) \in \Phi\}$ where $\Phi := \{y \in \mathbb{R} : \ y \leq 0\}$. Let d_S be the distance function. Then, for all $\zeta \in \partial^C d_S(x)$ there exists an $\alpha \in N^C_\Phi(h(x))$ such that

$$\zeta = \alpha \nabla h(x). \tag{6.1}$$

PROOF: By definition of S we have

$$d_S(x) = 0 \qquad \Longleftrightarrow \qquad h(x) \leq 0.$$

The set Φ is convex and thus $N^L_\Phi(y) = N^C_\Phi(y)$ for all $y \in \mathbb{R}$. If, for some $x \in \mathbb{R}^n$, we have $h(x) < 0$, then $N^C_\Phi(h(x)) = \{0\}$. If, however, $h(x) = 0$, then

$$\alpha \in N^C_\Phi(h(x)) \quad \Longrightarrow \quad \alpha \geq 0.$$

Recall that (see [13, 75] for example)

$$\partial^C d_S(x) \subset N^C_S(x).$$

Then

$$\zeta \in \partial^C d_S(x) \quad \Longrightarrow \quad \zeta \in N^C_S(x).$$

By (H4), if $\alpha \in \mathbb{R}$ such that $\alpha \geq 0$ and $\alpha \nabla h(x) = 0$, then $\alpha = 0$. It follows from Proposition 4.1 in [20] and (H4) that

$$\forall\ \zeta \in \partial^C d_S(x) \ \ \exists\, \alpha \in N^C_\Phi(h(x)) : \ \zeta = \alpha \nabla h(x),$$

which completes the proof. ∎

We now state the following *hypothetical* assumption:

(HH) Any strong minimum of (P) is also a strong minimum of the problem (Q):

$$(Q) \quad \begin{cases} \text{Minimize } l(x(a), x(b)) + K \displaystyle\int_a^b d_S(x(t))dt \\ \text{subject to} \\ \quad \dot{x}(t) = f(t, x(t), u(t)) \quad \text{a.e. } t \in [a, b], \\ \quad u(t) \in U(t) \quad \text{a.e. } t \in [a, b], \\ \quad (x(a), x(b)) \in E. \end{cases}$$

where $K > K_l$ and the set S is as defined in Lemma 6.1.1.

Problem (Q) is an optimal problem without state constraints. The state constraint $h(x(t)) \leq 0$ in (P) is incorporated in the cost function of (Q) via the integral of the distance function d_S.

We call (HH) a *hypothetical* assumption since what we would like to have is an assumption implying (HH). Let us see what we would need to guarantee (HH). It is a simple matter to see that (x^*, u^*) is an admissible solution to (Q). Moreover, any admissible process (z, v) for (P) is an admissible process for (Q). Suppose that (x^*, u^*) is not a strong minimum to (Q). Since $h(x^*(t)) \leq 0$ for all $t \in [a, b]$ we have $K \int_a^b d_S(x^*(t))dt = 0$. Consider an admissible process for (Q) (x', u') such that

$$l(x'(a), x'(b)) + K \int_a^b d_S(x'(t))dt < l(x^*(a), x^*(b)).$$

Set now

$$\rho = l(x^*(a), x^*(b)) - l(x'(a), x'(b)) - K \int_a^b d_S(x'(t))dt.$$

Then $\rho > 0$. Choose some $\delta \in (0, \frac{\rho}{2K})$. We have

$$0 < K\delta \leq \frac{\rho}{2} < \rho.$$

Consequently,

$$0 < l(x^*(a), x^*(b)) - K\delta - l(x'(a), x'(b)) - K \int_a^b d_S(x'(t))dt.$$

It follows that

$$l(x'(a), x'(b)) + K \int_a^b d_S(x'(t))dt < l(x^*(a), x^*(b)) - K\delta.$$

Suppose now that there exists an admissible process (z, v) for (P) such that

$$\max_{t\in[a,b]}\{|z(t) - x'(t)|\} \leq \frac{K}{2} \int_a^b d_S(x'(t))\, dt. \tag{6.2}$$

We know that

$$\Big|(z(a), z(b)) - (x'(a), x'(b))\Big| \leq 2 \max_{t\in[a,b]}\{|z(t) - x'(t)|\}$$

and, since l is Lipschitz, we have

$$l(z(a), z(b)) - l(x'(a), x'(b)) \leq K_l\Big|(z(a), z(b)) - (x'(a), x'(b))\Big|.$$

Thus, since $K > K_l$, we have

$$\begin{aligned} l(z(a),z(b)) - l(x'(a),x'(b)) &\leq K\Big|(z(a),z(b)) - (x'(a),x'(b))\Big| \\ &\leq K\int_a^b d_S(x'(t))dt \\ &< K\int_a^b d_S(x'(t))dt + K\delta \\ &< K\int_a^b d_S(x'(t))dt + \rho \\ &= K\int_a^b d_S(x'(t))dt + l(x^*(a),x^*(b)) - l(x'(a),x'(b)) - K\int_a^b d_S(x'(t))dt \\ &= l(x^*(a),x^*(b)) - l(x'(a),x'(b)) \end{aligned}$$

and we deduce that

$$l(z(a),z(b)) < l(x^*(a),x^*(b)).$$

This means that (x^*,u^*) is not optimal for (P), a contradiction. We summarize our findings:

Lemma 6.1.2 Let (x^*,u^*) be a strong minimizer for (P) and assume that (H1)-(H6) are satisfied. If, for any admissible process (x',u') of (Q) with $x'(t) \in x^*(t)+\varepsilon\bar{B}$, there exists an admissible process (z,v) of (P) with $z(t) \in x^*(t)+\varepsilon\bar{B}$ satisfying

$$\|z - x'\|_\infty \leq \frac{K}{2}\int_a^b d_S(x'(t))dt. \tag{6.3}$$

then (HH) holds.

Unfortunately, (6.3) is not satisfied in general. The existence of an admissible process (z,v) for (P) satisfying conditions somewhat similar to (6.3) has been vastly explored in the literature (see, for example, [38, 6, 7, 8]). However, no conditions involving $\int_a^b d_S(x'(t))dt$ are known to hold. Nevertheless as we can see next, if some conditions on the data of (P) would imply (6.3), they would be of use as we illustrate next.

A Hypothetical Theorem

Theorem 6.1.3 Let (x^*,u^*) be a strong minimum for (P). Assume that (H1)–(H6) and (HH) hold. Then there exists an absolutely continuous function p, a measurable function ξ and a scalar $\lambda \geq 0$

such that

(i) $||p||_\infty + \lambda > 0,$

(ii) $-\dot{p}(t) \in \partial_x^C \langle p(t), f(t, x^*(t), u^*(t))\rangle - \lambda\xi(t)\nabla h(x^*(t))$ a.e.,

(iii) $u \in U \implies \langle p(t), f(t, x^*(t), u)\rangle \leq \langle p(t), f(t, x^*(t), u^*(t))\rangle$ a.e.,

(iv) $(p(a), -p(b)) \in N_E^L(x^*(a), x^*(b)) + \lambda\partial^L l(x^*(a), x^*(b)),$

(v) $\xi(t) \geq 0$ and $\xi(t)h(x^*(t)) = 0$ a.e.

Notably, no measure is present in the above conditions.

PROOF: Let the problem (Q) first be translated into Mayer form as

$$(Q_M) \quad \begin{cases} \text{Minimize } l(x(a), x(b)) + y(b) \\ \text{subject to} \\ \dot{x}(t) = f(t, x(t), u(t)) \quad \text{a.e. } t \in [a, b], \\ \dot{y}(t) = Kd_S(x(t)) \quad \text{a.e. } t \in [a, b], \\ u(t) \in U(t) \quad \text{a.e. } t \in [a, b], \\ (x(a), x(b), y(a), y(b)) \in E \times \{0\} \times \mathbb{R}. \end{cases}$$

Applying the nonsmooth maximum principle without state constraints (see Theorem 2.5.4 or, alternatively, Theorem 6.2.1 in [75]) to (Q_M) we obtain the existence of absolutely continuous functions p_1, p_2 and a scalar $\lambda > 0$ such that

(a) $||p_1||_\infty + ||p_2||_\infty + \lambda > 0,$

(b) $(-\dot{p}_1(t), -\dot{p}_2(t)) \in \partial_{x,y}^C \langle (p_1(t), p_2(t)), (f(t, x^*(t), u^*(t)), Kd_S(x^*(t)))\rangle$ a.e.,

(c) $\langle (p_1(t), p_2(t)), (f(t, x^*(t), u^*(t)), Kd_S(x^*(t)))\rangle$
$= \max_{u\in U} \langle (p_1(t), p_2(t)), (f(t, x^*(t), u), Kd_S(x^*(t)))\rangle$ a.e.,

(d)
$$(p_1(a), -p_1(b), p_2(a), -p_2(b)) \in \lambda\partial^L_{x_1,x_2,y_1,y_2}\tilde{l}(x^*(a), x^*(b), y^*(a), y^*(b)) + N_E^L(x^*(a), x^*(b)) \times \mathbb{R} \times \{0\},$$

where $\tilde{l}(x_1, x_2, y_1, y_2) = l(x_1, x_2) + y_2$.

Since the scalar product $\langle (p_1(t), p_2(t)), [f(t, x^*(t), u^*(t)), Kd_S(x^*(t))] \rangle$ does not depend on y, we conclude from (b) that

$$\begin{aligned} -\dot{p}_1(t) &\in \partial_x^C \langle p_1(t), f(t, x^*(t), u^*(t)) \rangle + \partial_x^C \langle p_2(t), Kd_S(x^*(t)) \rangle \quad \text{a.e.} \\ \dot{p}_2(t) &= 0. \end{aligned} \tag{6.5}$$

Thus p_2 is a constant function. Lemma 6.1.1 and (6.5) yield

$$-\dot{p}_1(t) \in \partial_x^C \langle p_1(t), f(t, x^*(t), u^*(t)) \rangle + Kp_2\alpha(t)\nabla h(x^*(t)) \tag{6.6}$$

where $\alpha(t) \in N_\Phi^C (h(x^*(t))$ for a.e $t \in [a, b]$. Due to the properties of normal cones, the multifunction $t \to N_\Phi^C (h(x^*(t))$ is measurable, nonempty and closed, and thus has a measurable selection (see e.g. 3.1.1 in [13]). Thus the function $t \to \alpha(t)$ is measurable.

Since p_2 is constant and the cost function, defined as $\tilde{l}(x_1, x_2, y_1, y_2) = l(x_1, x_2) + y_2$, is independent of y_1 we deduce from the tranversality condition (d) that

$$-p_2(b) = \lambda \partial_{y_2}^L [l(x^*(a), x^*(b)) + y^*(b)] = \lambda$$

and, subsequently, $p_2 \equiv -\lambda$. Based on (6.6), we write the Euler-Lagrange inclusion, changing the notation of p_1 into p and setting $\xi(t) := \frac{1}{K}\alpha(t)$, as

$$-\dot{p}_1(t) \in \partial_x^C \langle p_1(t), f(t, x^*(t), u^*(t)) \rangle - \lambda\xi(t)\nabla h(x^*(t)) \quad \text{a.e.}, \tag{6.7}$$

which corresponds to (ii).

Recall that the function α has the property

$$\alpha(t) \begin{cases} = 0, & \text{if } h(x^*(t)) < 0, \\ \geq 0, & \text{if } h(x^*(t)) = 0. \end{cases}$$

This property leads to the complementary slackness condition (v).

The Weierstrass condition (c), after calculating the scalar products, reduces to

$$\langle p(t), f(t, x^*(t), u) \rangle \leq \langle p(t), f(t, x^*(t), u^*(t)) \rangle \quad \text{a.e. for } u \in U,$$

which is exactly (iii) of our theorem. Finally, taking into account that $p_2 \equiv -\lambda$, if $\lambda = 0$ then the condition (a) reads as $\|p_1\|_\infty > 0$ and thus implies (i). If $\lambda > 0$ then the conditions (a) and (i) are clearly satisfied. ∎

6.2 A First Order Problem

Let us now turn to autonomous problems of the form

$$(\text{FO}) \quad \begin{cases} \text{Minimize } \displaystyle\int_a^b \left(\langle c, x\rangle + u^2\right)\, dt \\ \text{subject to} \\ \quad \dot{x}(t) = f(x(t)) + g(x(t))u(t) \text{ a.e.}t, \\ \quad h(x(t)) \le 0 \quad \forall t, \\ \quad u(t) \in U \quad \text{a.e.}t, \\ \quad x(a) = x_a, \end{cases}$$

where $c \in \mathbb{R}^n$, u is a scalar, U is a compact set in $\mathbb{R}$ and $\nabla h(x) \neq 0$ whenever $h(x) = 0$. Here, as before, we assume that h is a scalar valued function. Let us briefly review some concepts on the state constraint appearing in (FO) that will be important in our setting. A boundary interval for the state constraint along a trajectory x of (FO) is an interval $[t_0^b, t_1^b] \subset [a, b]$ if it is the maximal interval where $h(x(t)) = 0 \quad \forall t \in [t_0^b, t_1^b]$. The point t_0^b and t_1^b are called *entry point* and *exit point*, respectively. Any interval $I \subset [a, b]$ is an *interior interval* if $h(x(t) < 0 \quad \forall t \in I$. A point $\sigma \in [a, b]$ is a *contact point* for x if it is an isolated point such that $h(x(\sigma)) = 0$.

Problem (FO) has one state constraint. Let (x^*, u^*) be local strong minimum for (FO) and assume that our conditions (H1)–(H6) are satisfied. Theorem 9.3.1 in [75] asserts that there exist an absolutely continuous function p, a scalar λ, a measure $\mu \in C^{\oplus}([a, b])$ such that

(i) $(p, \lambda, \mu) \neq (0, 0, 0)$;

(ii) $-\dot{p}(t) = f_x^T(x^*(t))q(t) + u^*(t)g_x^T(x^*(t))q(t) - \lambda c$;

(iii) for all $u \in U$, $\langle g(x^*(t))u^*(t), q(t)\rangle - \lambda(u^*)^2(t) \ge \langle g(x^*(t))u, q(t)\rangle - \lambda u^2$,

(iv) $-q(b) = 0$;

(v) $supp\{\mu\} \subset \{t : \ h(x^*(t)) = 0\}$.

where

$$q(t) = p(t) + \int_{[a,t)} \nabla h(x^*(s))\, \mu(ds),$$
$$q(b) = p(b) + \int_{[a,b]} \nabla h(x^*(s))\, \mu(ds).$$

If the above conditions hold with $\lambda = 1$, then we say that the problem is *normal*.

In [73] (see also [38] and the references within), conditions are derived for the problems of the type of (FO) to guarantee that the measure associated with the state constraint in the normal form of the maximum principle is *regular* (in the sense that it is absolutely continuous with respect to the Lebesgue measure in the interior of the interval $[a, b]$). One might think that conditions imposed to assert regularity of the adjoint variable, if satisfied, would also imply that our hypothetical assumption (HH) holds. However, we need to keep in mind that neither [73] nor [38] provide us with information about the possible behaviour at the points $t = a$ or $t = b$. If we knew a priori that $h(x^*(a)) < 0$ and $h(x^*(b)) < 0$, then, the regularity of the measure would follow. But no such guarantee exists as we illustrate next with a simple problem with one state constraint of first order recovered from [9].

A Case Study with State Constraints

Problem (P_S) of Chapter 5 is a particular instance of problem (FO). Our findings in Section 5.4 show that (P_S) is normal and that the adjoint variable is regular. However, our numerical findings of (P_S) also show that the state constraint is active at the end point and that the measure has an atom at $t = T$. Thus, the SEIR problem serves also as an counter-example that the assumption (HH) is not generally satisfied.

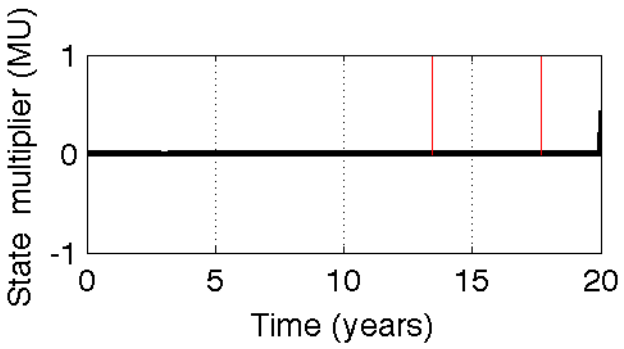


Figure 6.1: The state multiplier exibiting a jump at the end point

Once more, the pathological aspect is the fact that the state constraint is active at the end point.

6.3 Conclusions

We showed here that exact penalization for state constrained problems holds for problems satisfying (HH). Since this assumption is difficult (if not impossible) to verify in particular situations, we went a step further and defined a sufficient condition for (HH) to hold. The exact penalization approach outlined here would guarantee the absolute continuity of the multiplier μ in $[0, T]$. However, we showed via a very simple example with a first order state constraint that exact penalization scheme is not satisfied. We conclude that identification of classes of problems with an absolutely continuous multiplier μ is a difficult subject.

Chapter 7

The SEIR Problem with Mixed Constraints and L^1 Cost

In this chapter we introduce an L^1 type cost linear with respect to the control variable of the SEIR optimal control problem of Chapter 4. Such a cost functional is considered more appropriate for problems with a biological or medical background than an L^2 cost. If the control function stands for a medication or a vaccination policy and the admissible control values range on a small scale, a L^1 cost gives the control values between 0 and 1 a greater impact than a control-quadratic cost. Besides, resulting bang-bang controls are often easier to implement in the biomedical praxis.

We introduce a pointwise limitation on the stock of available vaccine which results in a mixed state-control constraint where the control appears linearly. In this case the augmented Hamiltonian function is linear with respect to the control and the necessary conditions of optimality show that any optimal control must be a concatenation of bang-bang and singular arcs. Although no singular arcs appear in our problem, in general, singular control arcs can be determined via the switching function which we introduce to the given problem. The numerical solution is obtained in a similar manner as in Chapter 5 via discretizing the control problem and applying nonlinear programming with ICLOCS and IPOPT. Although we do not show that the numerical solution is indeed a (local) optimum, we do however validate our findings. Using the Lagrange multipliers provided by the optimization solver, we can validate our numerical solution by showing that it satisfies precisely the necessary condition of optimality.

7.1 The Optimal Control Problem with Mixed Constraints

We consider the SEIR optimal control with a mixed control-state constraint

$$(P_1)\quad \begin{cases} \text{Minimize} \quad J_1(x,u) = \displaystyle\int_0^T (AI(t) + Bu(t))\ dt \\ \text{subject to} \\ \quad \dot{S}(t) = bN(t) - dS(t) - cS(t)I(t) - u(t)S(t), \\ \quad \dot{E}(t) = cS(t)I(t) - (e+d)E(t), \\ \quad \dot{I}(t) = eE(t) - (g+a+d)I(t), \\ \quad \dot{N}(t) = (b-d)N(t) - aI(t), \\ \quad u(t)S(t) \le V_0, \\ \quad u(t) \in [0,1] \quad \text{for a.e. } t \in [0,T], \\ \quad S(0) = S_0,\ E(0) = E_0,\ I(0) = I_0,\ N(0) = N_0. \end{cases}$$

which differs from problem (P_S) by the presence of an L^1 cost with different values for A, B than those in Section 5.3 and the mixed constraint $u(t)S(t) \le V_0$ instead of $S(t) \le S_{max}$.

To simplify the analysis of the necessary optimality conditions of the control problem (P_1), it is convenient to rewrite it in the form of a general optimal control problem with a mixed control-state constraint:

$$(\mathrm{P}_{\mathrm{mixed}})\quad \begin{cases} \text{Minimize} \displaystyle\int_0^T L(x(t),u(t))\,dt \\ \text{subject to} \\ \quad \dot{x}(t) = f(x(t)) + g(x(t))u(t) \text{ a.e. } t \in [0,T], \\ \quad m(x(t),u(t)) \le 0 \text{ a.e. } t \in [0,T], \\ \quad u(t) \in [0,1] \text{ a.e. } t \in [0,T], \\ \quad x(0) = x_0, \\ \quad x(T) \in \mathbb{R}^n, \end{cases}$$

where

$$\begin{aligned} &x = (S,E,I,N), && L(x,u) = AI + Bu = L_1(x) + L_2(u), \\ &f(x) = f_1(x) + A_1 x, && f_1(x) = c(-SI, SI, 0, 0)^T, \\ &g(x) = (-S,0,0,0)^T, && m(x,u) = uS - V_0, \end{aligned}$$

and A_1 as previously defined in Section 5.3. The initial condition x_0 and parameters will be assumed as in Table 4.1. The differential equation $\dot{x}(t) = f(x(t)) + g(x(t))u(t)$ is affine in the control and is nonlinear in the state x due to the term $f_1(x)$. Note that the mixed control-state constraint satisfies

the standard *regularity condition*

$$m_u(x(t), u(t)) = S(t) \neq 0 \quad \forall\, t \in [0, T] \;\text{ with }\; u(t)S(t) = V_0\,. \tag{7.1}$$

7.2 Discussion of Necessary Conditions for (P_1)

Let (x^*, u^*) be a minimizer for our problem (P_1) (or (P_{mixed})). In the following, we shall evaluate the necessary optimality condition of the *Maximum Principle*. Since we are maximizing $-J_1(x, u)$, the standard Hamiltonian function is given by

$$H(x, p, u) = -\lambda L(x, u) + \langle p, f(x) + g(x)u\rangle, \quad \lambda \in \mathbb{R},$$

where $p = (p_s, p_e, p_i, p_n) \in \mathbb{R}^4$ denotes the adjoint variable. In the *augmented* Hamiltonian, the mixed constraint $m(x, u) \geq 0$ is adjoined by a multiplier $q \in \mathbb{R}$ to the Hamiltonian:

$$\mathcal{H}(x, p, q, u) = H(x, p, u) - q\, m(x, u).$$

Here, the minus sign is due to the fact that the *Maximum Principle* assumes that the control-state constraint is written in the form $-m(x, u) \geq 0$. In view of the regularity condition (7.1), Theorem 7.1 in [20] (cf. also [43, 57]) asserts the existence of a scalar $\lambda \geq 0$, an absolutely continuous function $p : [0, T] \to \mathbb{R}^4$ and an integrable function $q : [0, T] \to \mathbb{R}$ such that the following conditions are satisfied almost everywhere:

(i)

$$\|p\|_\infty + \lambda > 0, \tag{7.2}$$

(ii) (adjoint equation and transversality condition)

$$\begin{aligned} -\dot{p}(t) &= \mathcal{H}_x(x^*(t), p, q, u^*(t)) \\ &= -\lambda L_x(x^*(t), u^*(t)) + \langle p(t), f_x(x^*(t)) + g_x(x^*(t))u^*(t)\rangle - \langle q(t), m_x(x^*(t), u^*(t))\rangle\,, \\ -p(T) &= (0, 0, 0, 0), \end{aligned} \tag{7.3}$$

(iii) (maximum condition for Hamiltonian H)

$$H(x^*(t), p(t), u^*(t)) = \max_u \{\, H(x^*(t), p(t), u) \mid 0 \le u \le 1,\ m(x_*(t), u) \le 0 \,\}, \tag{7.4}$$

(iv) (local maximum condition for augmented Hamiltonian $\mathcal{H}$)

$$\begin{aligned} \mu(t) &= \mathcal{H}_u\left(x^*(t), p(t), q(t), u^*(t)\right) \\ &= -L_u(x^*(t), u^*(t)) + \langle p(t), g(x^*(t)\rangle - q(t)\, m_u(x^*(t), u^*(t)) \ \in N_{[0,1]}(u^*(t)), \end{aligned} \tag{7.5}$$

(v) (complementarity condition)

$$q(t)\, m(x^*(t), u^*(t)) = q(t)\ [u^*(t) S_*(t) - V_0] = 0 \quad \text{and} \quad q(t) \ge 0. \tag{7.6}$$

In (7.5), $N_{[0,1]}(u^*(t)) = \{0\}$ when $u^*(t) \in (0,1)$. Since the terminal state $x(T)$ is free, it is easy to prove that the above necessary conditions hold with $\lambda = 1$; for a complete discussion see [9]. Hence, our problem is *normal*. We can further prove the existence of a constant K_q^1 such that

$$|q(t)| \le K_q^1 |p(t)| \tag{7.7}$$

for almost every $t \in [0, T]$ (see [20]).

Now we want to extract information from the conclusions (7.2)–(7.6) with $\lambda = 1$ that later will be used to validate our numerical solution. The adjoint equations in (7.3) for the adjoint variable $p = (p_s, p_e, p_i, p_n)$ are explicitly given by

$$-\dot{p}_s(t) = -(d + cI_*(t) + u^*(t))p_s(t) + cI_*(t)p_e(t) - u^*(t)q(t), \tag{7.8}$$

$$-\dot{p}_e(t) = -(e + d)p_e(t) + ep_i(t), \tag{7.9}$$

$$-\dot{p}_i(t) = -cS_*(t)p_s(t) + cS_*(t)p_e(t) - (g + a + d)p_i(t) - ap_n(t) - A, \tag{7.10}$$

$$-\dot{p}_n(t) = bp_s(t) + (b - d)p_n(t). \tag{7.11}$$

Next, we evaluate the maximum condition (7.4) for the Hamiltonian H. We define the *switching function* ϕ by

$$\phi(x, p) = H_u(x, u, p) = -B - p_s\, S, \qquad \phi(t) = \phi(x(t), p(t)) \tag{7.12}$$

and see that the condition (7.4) is equivalent to the maximum condition

$$\phi(t)u^*(t) = \max_u \{\, \phi(t)u \mid 0 \le u \le 1,\, u\, S_*(t) \le V_0 \,\}. \tag{7.13}$$

This yields the control law

$$u^*(t) = \begin{cases} \min\left\{1, \dfrac{V_0}{S_*(t)}\right\} & , \quad \text{if} \quad \phi(t) > 0 \\ 0 & , \quad \text{if} \quad \phi(t) < 0. \end{cases} \tag{7.14}$$

Any isolated zero of the switching function $\phi(t)$ yields a switch of the control from $\min\{1, V_0/S_*(t)\}$ to 0 or vice versa. If, however, $\phi(t) = 0$ holds on an interval $[t_1, t_2] \subset [0, T]$, then we have a *singular control*. We do not enter here into a detailed discussion of singular controls, since they never appeared in our computations. Moreover, our computations show that $0 < u^*(t) < 1$ holds along a boundary arc of the mixed constraint $uS \le V_0$, i.e., whenever $u^*(t) = V_0/S_*(t)$. Hence, the control is determined by

$$u^*(t) = \begin{cases} V_0/S_*(t), & \text{if} \quad \phi(t) > 0 \\ 0, & \text{if} \quad \phi(t) < 0. \end{cases} \tag{7.15}$$

Due to $0 < u^*(t) < 1$ the multiplier $\mu(t)$ in (7.5) vanishes which yields the relation

$$0 = \mu(t) = \mathcal{H}_u(x^*(t), p(t), q(t), u^*(t)) = -B - p_s(t)S_*(t) - q(t)S_*(t).$$

This allows us to compute the multiplier $q(t)$ for which we get in view of the complementarity condition (7.6)

$$q(t) = \begin{cases} -\dfrac{B}{S_*(t)} - p_s(t) = \phi(t)/S_*(t), & \text{if} \quad u^*(t) = V_0/S_*(t), \\ 0, & \text{if} \quad u_*(t) < V_0/S_*(t)\ . \end{cases} \tag{7.16}$$

7.3 Numerical Results

We keep the parameter values and the initial values $S(0), E(0), I(0), N(0)$ as presented in Table 4.1 except for the new values $A = 5$ and $B = 10$. We also rely on the same software and the choice of T, N and the acceptable converegence tolerance ε_{rel} as presented in Section 4.3.

For the mixed constraint $u(t)S(t) \leq V_0 = 125$ we find the optimal control

$$u^*(t) = \begin{cases} 125/S_*(t) & \text{for} \quad 0 \leq t \leq t_1, \\ 0 & \text{for} \quad t_1 < t \leq T = 20. \end{cases} \tag{7.17}$$

This shows that the constraint itself when expressed as the *new control variable* $v = uS$ is a bang-bang control with only one switch at t_1; cf. [56]. We obtain the numerical results

$$\begin{aligned} &J_1(x,u) = 1692.2, \quad t_1 = 17.89, \\ &S(T) = 1723.8, \quad E(T) = 7.7030, \quad I(T) = 4.7038, \quad N(T) = 1824.2. \end{aligned}$$

The total amount of used vaccines is $W(T) = 2235.8$. The optimal trajectories and optimal control are presented in Figure 7.1.

The adjoint variables are displayed in Figure 7.2. It can be seen in Figure 7.3 that the switching function $\phi(t)$ satisfies exactly the control law (7.14) while Figure 7.4 shows that the multiplier $q(t)$ obeys the multiplier rule (7.16). Thus the computed solution is precisely identical to the analytic solution.

7.4 Conclusion

We considered an optimal control problem with mixed constraints and L^1 cost for a SEIR epidemic model of human infectious diseases. In this optimal control problem the control appears linearly. We discussed the necessary conditions of the Maximum Principle and obtained explicit formulas for the switching function and the multiplier associated with the mixed constraint in terms of state and adjoint variables. We have seen that singular controls have never appeared in this problem.

Since the numerical approach furnishes as well the adjoint variables, we could verify that the computed solution satisfies the necessary optimality conditions precisely. The numerical verification of second-order sufficient conditions using the methods in [63, 56] remains a future work.

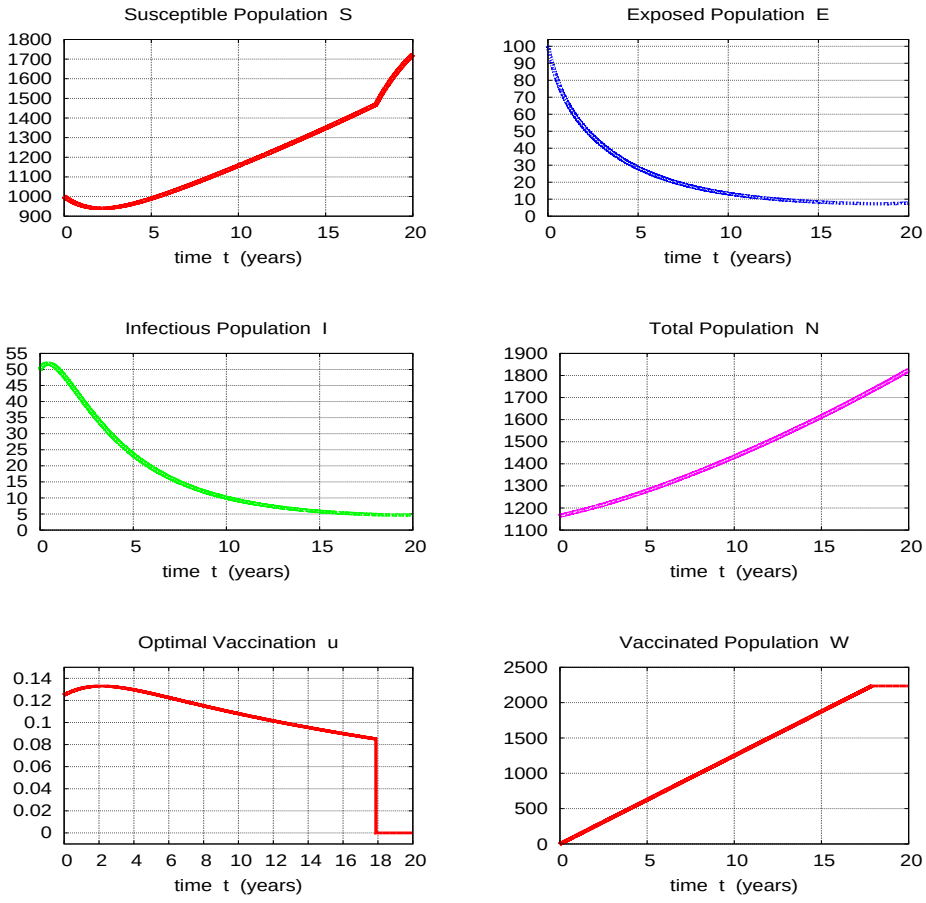


Figure 7.1: Optimal trajectories and control (vaccination) for mixed constraint $uS \leq 125$.
Top row: (*left*) susceptible population S, (*right*) exposed population E.
Middle row: (*left*) infectious population I, (*right*) total population N.
Bottom row: (*left*) vaccination (control) u, (*right*) vaccinated population W.

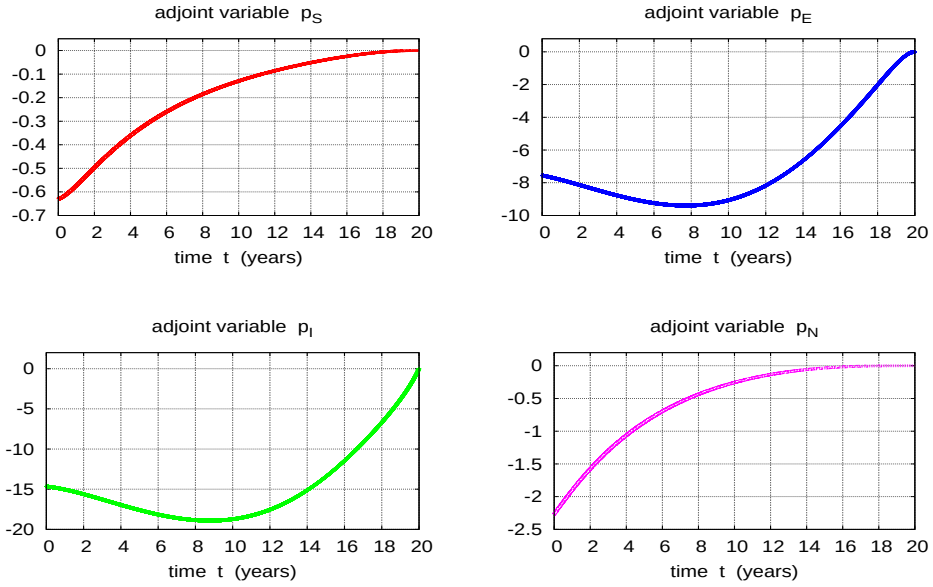


Figure 7.2: Adjoint variables, multiplier and switching function for mixed constraint $u(t)S(t) \leq 125$.
Top row: (*left*) adjoint variable p_S, (*right*) adjoint variable p_E.
Bottom row: (*left*) adjoint variable p_I, (*right*) adjoint variable p_N.

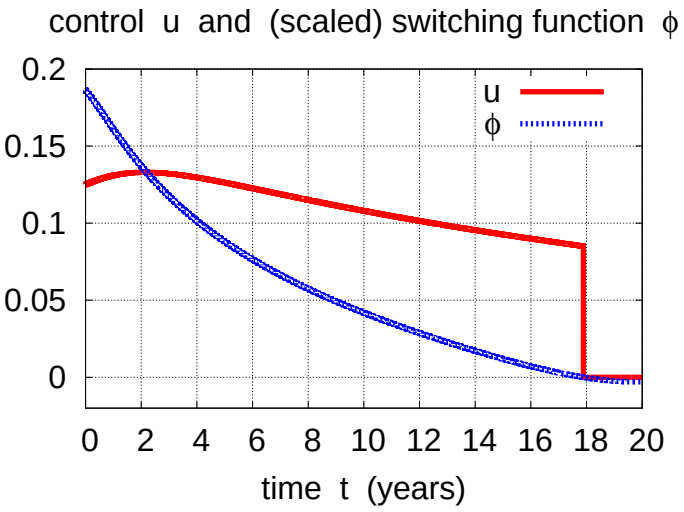


Figure 7.3: Control u and (scaled) switching function $\phi(t)$ satisfying the control law (7.14).

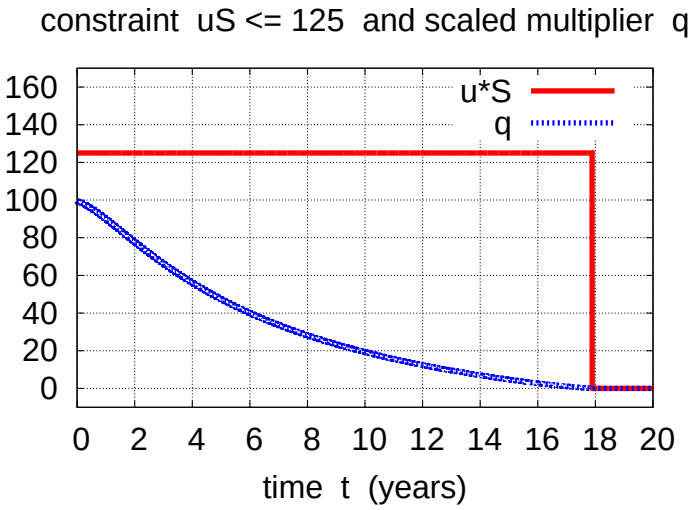


Figure 7.4: Constrained function uS and multiplier q satisfying (7.16).

Chapter 8

Optimal Control Problems with Differential Algebraic Equations

The main contribution presented in this chapter is the formulation of necessary conditions for the existence of an optimal control (Maximum Principle) for optimal control problems involving Differential Algebraic Equations (DAE). Optimality conditions for DAE control problems are a challenging subject (see, for example, [23, 31, 71, 39, 50]). Our necessary conditions, based upon the results of [20], treat DAE problems as mixed constrained problems and include the nonsmooth case. A central point is that we do not need to apply the computationally expensive implicit function theorem. This work has been presented at the *SADCO Doctoral Days 2012* in Paris and *MTNS 2012*, Melbourne (see [46]).

8.1 DAE control problems

DAE control problems appear often in robotics, economics and process systems engineering. For a typical *differential algebraic equation* in the *semi-explicit form*

$$(DAE) \qquad \begin{cases} \dot{x}(t) = f(t, x(t), y(t), u(t)) \text{ a.e.} \\ 0 = g(t, x(t), y(t), u(t)) \text{ a.e.} \end{cases}$$

and its state vector (x, y) we distinguish between the "slow" state variable x, where the derivatives are given, and the "fast" state variable y which is constrained merely by algebraic equations and can respond rapidly to changes in control u. Formally, x and y are referred to as *differential* and *algebraic* variables, respectively.

A pair (x, u) is called a *solution* to (DAE) if, for a given control function $u : [a, b] \to \mathbb{R}^k$, the equations

$$\begin{cases} x(t) = x(t_0) + \displaystyle\int_{t_0}^{t} f(s, x(s), y(s), u(s))\, ds \\ 0 = g(t, x(t), y(t), u(t)) \end{cases}$$

are satisfied for $t \in [a, b]$. Recall that u is assumed to be merely a measurable function.

Our standard control problem involving DAE is

$$(P_{DAE}) \qquad \begin{cases} \text{Minimize } l(x(a), x(b)) \\ \text{subject to} \\ \quad \dot{x}(t) = f(t, x(t), y(t), u(t)) \text{ a.e.} \\ \quad 0 = g(t, x(t), y(t), u(t)) \text{ a.e.} \\ \quad u(t) \in U \text{ a.e.} \\ \quad (x(a), x(b)) \in E \end{cases}$$

where $l : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, $f : [a, b] \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^k \to \mathbb{R}^n$, $g : [a, b] \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^k \to \mathbb{R}^m$, $U \subset \mathbb{R}^k$ is compact and $E \subset \mathbb{R}^n \times \mathbb{R}^n$ is a closed set.

For many specific problems involving DAE's necessary optimality conditions have been derived when the fast variable, y, is treated as an extra control. Such problems can be treated as mixed state-control constrained problems. On the other hand, the variable y can be considered as part of the state variable $z = (x, y)$. This latter approach lets us view both differential and algebraic equations in (DAE) as a single equation $h(t, z, u, \dot{z}) = 0$ where

$$h(t, z, u, \dot{z}) = E\dot{z}(t) - \begin{pmatrix} f(t, z(t), u(t)) \\ -g(t, z(t), u(t)) \end{pmatrix} = 0,$$

and $E = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$ with I being the identity matrix. We say that

$$h(t, z, u, \dot{z}) = 0$$

is an equivalent formulation for (DAE) in the *implicit form*.

In studying DAE systems it is helpful to divide them into classes. One such classification as far as necessary conditions of optimality are concerned is the one related to the differentiation index. Rather informally, one says that *(DAE)* is of index $(K + 1)$ if we need to differentiate the algebraic equation K times to get y as a function of x, u and derivatives of u To illustrate this, we consider

two cases.

Consider the system (DAE) where g is continuously differentiable with respect to (x, y, u) and, for almost every t and all (x, y, u), $g_y(t, x, y, u)$ is nonsingular and bounded. Under this condition, (DAE) is an index 1 system. Then, appealing to the implicit function theorem, we can solve locally $g(t, x, y, u) = 0$ to obtain $y = \varphi(t, x, u)$, φ being the implicit function. Then, at least locally, we can replace (DAE) by

$$\dot{x}(t) = f(t, x(t), \varphi(t, x(t), u(t)), u(t)).$$

As an example of a DAE of index higher than 1 consider

$$\begin{cases} \dot{x}(t) = f(t, x(t), y(t), u(t)), \\ 0 = g(t, x(t)) \end{cases}$$

and assume that g is continuously differentiable. Differentiating the algebraic equation again with respect to t (assuming the derivatives all exist and are continuous) we can write

$$0 = \frac{d}{dt} g(t, x(t)) = \frac{\partial g(t, x(t))}{\partial t} + \frac{\partial g(t, x(t))}{\partial x} f(t, x(t), y(t), u(t)) =: G(t, x, y, u)$$

If G_y is nonsingular then the DAE system is of index 2. Otherwise we may have to differentiate the last equation again and again.

8.2 Index One: Nonsmooth Case

We remind in the current context of two definitions previously introduced in Section 2.1. We call (x, y, u) an *admissible process* to the problem (P_{DAE}) if the triple satisfies (DAE) whereas $x : [a, b] \to \mathbb{R}^n$ is an absolutely continuous function, $y : [a, b] \to \mathbb{R}^m$, $u : [a, b] \to \mathbb{R}^k$ are measurable functions such that

$$u(t) \in U \quad \text{a.e. } t \in [a, b],$$

where U is a compact set. The process (x^*, y^*, u^*) is a $W^{1,1}$ *local minimum* for the optimal control problem (P_{DAE}) if, for some $\varepsilon > 0$, it minimizes the cost over all other admissible processes (x, y, u) such that

$$\|x - x^*\|_\infty \leq \varepsilon \text{ and } \int_a^b |\dot{x}(t) - \dot{x}^*(t)| \, dt \leq \varepsilon.$$

The results of [20] are explored in what follows:

We treat the variable y as a control and operate with the following *basic hypotheses*: the function l is locally Lipschitz, E is a closed set, U is compact and $(t,(x,y,u)) \to (f(t,(x,y,u)), g(t,(x,y,u)))$ is a $\mathcal{L}\times\mathcal{B}$-measurable function (w.r.t. σ-field generated by the product of $\mathcal{L}$-measurable and $\mathcal{B}$-measurable subsets in $\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^k$). We make use of the sets

$$S(t) := \{(x,y,u) \,:\, g(t,x,y,u)=0,\, u\in U\}\,, \tag{8.1}$$

and

$$S^*_\varepsilon(t) := \{(x,y,u)\in S(t) \,:\, |x(t)-x^*(t)|\le\varepsilon\}\,. \tag{8.2}$$

Among other conditions, we require *Lipschitz continuity* for f and g.
Consider a function $\psi : [a,b]\times\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^k\to\mathbb{R}^n$ which stands representatively for either f or g. The condition

($\mathbf{L}^*$) $\exists\, k_\psi>0$ such that for (almost) every $t\in[a,b]$ and all (x_i,y_i,u_i) with $|x_i(t)-x^*(t)|\le\varepsilon$:

$$|\psi(t,x_1,y_1,u_1)-\psi(t,x_2,y_2,u_2)| \le k_\psi\left(|x_1-x_2|+|y_1-y_2|+|u_1-u_2|\right)$$

shall hold for both $\psi\equiv f$ and $\psi\equiv g$. Besides, we assume the following *calibrated constraint qualification* for the algebraic equation and the set control constraints S^*_ε:

(**A1**) $\exists\, M>0$ such that for almost every $t\in[a,b]$, all $(x,y,u)\in S^*_\varepsilon(t)$, all $\lambda\in\mathbb{R}^m$, all $\xi\in N^L_U(u)$

$$(\alpha,\beta_1,\beta_2-\xi)\in\partial^L_{x,y,u}\langle\lambda,g(t,x,y,u)\rangle \quad\Longrightarrow\quad |\lambda|\le M\,|(\beta_1,\beta_2)|\,.$$

Theorem 8.2.1 (Nonsmooth maximum principle) Suppose (x^*,y^*,u^*) is a $W^{1,1}$ local minimizer for (P_{DAE}). If the basic hypotheses are valid, f and g satisfy ($\mathbf{L}^*$), and (**A1**) holds, then there exists $p\in W^{1,1}([a,b];\mathbb{R}^n)$ and a scalar $\lambda_0\ge 0$ satisfying
the *nontriviality condition*:

$$\|p\|_\infty+\lambda_0>0,$$

the *Euler adjoint inclusion*: for almost every $t\in[a,b]$

$$(-\dot p(t),0,0)\in\partial^C_{x,y,u}\langle p(t),f(t,x^*(t),y^*(t),u^*(t))\rangle - N^C_{S(t)}(x^*(t),y^*(t),u^*(t)),$$

the global *Weierstrass condition*: for almost every $t\in[a,b]$ and all $(x^*(t),y,u)\in S(t)$

$$\langle p(t),f(t,x^*(t),y,u)\rangle \le \langle p(t),f(t,x^*(t),y^*(t),u^*(t))\rangle\,,$$

and the *transversality condition*:

$$(p(a), -p(b)) \in N_E^L(x^*(a), x^*(b)) + \lambda_0 \partial^L l(x^*(a), x^*(b))$$

PROOF: This is a direct application of Theorem 7.1 in [20] when the control variable is considered to be $v = (y, u)$. ∎

Remark 8.2.2 It is clear from the proof that the result holds when y is seen as a component of the control variable and not as a state.

It may be difficult to apply Theorem 8.2.1 in the praxis because the Euler adjoint inclusion relies on the set $N^C_{S(t)}(x^*(t), y^*(t), u^*(t))$ which is large and unbounded, and thus hard to handle. However, we can give up the troublesome normal cone and simplify the Euler adjoint inclusion if one assumes some differentiability of the function g, as we show in the next section.

8.3 Index One: Differential Case

Let us suppose that the function g is strictly differentiable along the optimal solution. Then the Euler adjoint equation can be written in terms of $\nabla_{x,y,u} g$, as we will see:

Proposition 8.3.1 Suppose (x^*, y^*, u^*) is a $W^{1,1}$ local minimizer for (P_{DAE}). Assume further, as in the described nonsmooth case, the basic hypotheses are valid, (**A1**) holds, ($\mathbf{L}^*$) is satisfied and $(x, y, u) \mapsto g(t, x, y, u)$ also is strictly differentiable at $(x^*(t), y^*(t), u^*(t))$ for almost every t.

Then, for any measurable function $\chi : [a, b] \to \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^k$ such that

$$\chi(t) \in N^C_{S(t)}(x^*(t), y^*(t), u^*(t)) \quad \text{a.e.}$$

there exist measurable functions $\lambda : [a, b] \to \mathbb{R}^m$, $\xi : [a, b] \to \mathbb{R}^k$ such that $\xi(t) \in N^C_U(u^*(t))$ and

$$\chi(t) = g_{x,y,u}(t, x^*(t), y^*(t), u^*(t))^T \lambda(t) + (0, 0, \xi(t)).$$

PROOF: Let Φ be a closed subset of $\mathbb{R}^q$ and let $\phi : \mathbb{R}^k \to \mathbb{R}^q$ be a locally Lipschitz continuous function in a neighbourhood of w such that $\phi(w) \in \Phi$. We will come back to the matter of the function ϕ shortly. Define the set $\mathcal{S} := \{w := (x, y, u) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^k \ : \ \phi(w) \in \Phi\}$. If the *constraint qualification* condition:

(CQ)

$$\eta \in N^L_\Phi(\phi(w^*)), \qquad 0 \in \partial_L \langle \eta, \phi \rangle (w^*) \qquad \Longrightarrow \qquad \eta = 0$$

is satisfied, then Proposition 4.1 in [20] implies:

> If ϕ is strictly differentiable at w^* and if $\zeta \in N_{\mathcal{S}}^C(w^*)$, then there exists $\eta \in N_{\Phi}^C(\phi(w^*))$ such that $\zeta = \nabla \langle \eta, \phi \rangle (w^*)$.

We now consider

$$\Phi := \{0\} \times U \subset \mathbb{R}^n \times \mathbb{R}^k \qquad \text{and} \qquad \phi(w) := [g(t,x,y,u), u],$$

where $w = (x,y,u) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^k$.

Let us fix a $t \in [a,b]$ and verify (CQ) for this t with the help of (A1). Assume that

$$\eta := (\lambda, \xi) \in N_{\Phi}^L(\phi(w^*(t))) = N_{\{0\}}^L\left(g(t, x^*(t), y^*(t), u^*(t))\right) \times N_U^L\left(u^*(t)\right),$$

or, equivalently, that $\lambda \in \mathbb{R}^m$, $\xi \in N_U^L(u^*(t))$. Note that the function ϕ is strictly differentiable since its two components are strictly differentiable at $(t, x^*(t), y^*(t), u^*(t))$. Assume further that $0 \in \partial_L \langle (\lambda, \xi), \phi(w^*(t)) \rangle$ which is equivalent to

$$\begin{aligned} (0,0,0) &= \nabla_{x,y,u} \langle (\lambda, \xi), \phi(w^*(t)) \rangle \\ &= \nabla_{x,y,u} \langle \lambda, g(t, x^*(t), y^*(t), u^*(t)) \rangle + (0,0,\xi), \end{aligned} \tag{8.4}$$

For convenience, we rewrite (8.4) as

$$(0,0,-\xi) = \nabla_{x,y,u} \langle \lambda, g(t, x^*(t), y^*(t), u^*(t)) \rangle \tag{8.5}$$

This allows us to invoke (A1). For $\alpha := 0$, $\beta_1 = 0$, $\beta_2 = 0$, (A1) yields with some $M > 0$ that

$$|\lambda| \leq M\,|(\beta_1, \beta_2)| = 0$$

and, thus, $\lambda = 0$. We have shown that (A1) implies (CQ) for $w^* = (x^*, y^*, u^*)$.

Let us return to Proposition 4.1 in [20]. For almost every $t \in [a,b]$ and every triple

$$(\zeta_1, \zeta_2, \zeta_3) \in N_{\mathcal{S}}^C(x^*(t), y^*(t), u^*(t))$$

there exists a pair

$$(\lambda, \xi) \in N_{\Phi}^C(\phi(w^*(t))) = N_{\{0\}}^C\left(g(t, x^*(t), y^*(t), u^*(t))\right) \times N_U^C\left(u^*(t)\right)$$

such that

$$\begin{aligned}(\zeta_1,\zeta_2,\zeta_3) &= \nabla_{x,y,u}\,\langle(\lambda,\xi),[g(t,x^*(t),y^*(t),u^*(t)),u^*(t)]\rangle\\ &= \nabla_{x,y,u}\,\langle\lambda,g(t,x^*(t),y^*(t),u^*(t))\rangle+(0,0,\xi).\end{aligned}\tag{8.7}$$

Since the normal cone is a closed and nonempty set for all $t\in[a,b]$, it admits a measurable selection and we may write (8.5) as

$$\chi(t):=(\zeta_1(t),\zeta_2(t),\zeta_3(t))=\nabla_{x,y,u}g(t,x^*(t),y^*(t),u^*(t)^T\lambda(t)+(0,0,\xi(t)).$$

where

$$(\zeta_1(t),\zeta_2(t),\zeta_3(t))\in N^C_{S(t)}(x^*(t),y^*(t),u^*(t))\ \text{ a.e.}$$

This proves the claim. ∎

Corollary 8.3.2 If besides the hypotheses of Theorem 8.2.1 the function g is strictly differentiable at (x^*,y^*,u^*), then the Euler adjoint inclusion is replaced by

$$(-\dot p(t),0,\xi(t))\in\partial^C_{x,y,u}\,\langle p(t),f(t,x^*(t),y^*(t),u^*(t))\rangle-\nabla g_{x,y,u}(t,x^*(t),y^*(t),u^*(t))^T\lambda(t),$$

where $\xi(t)\in N^C_U(u^*(t))$ almost everywhere.

In the literature the function $(x,y,u)\mapsto g(x,y,u)$ is assumed C^1. Next we see how the C^1 property relates to the well known smoothness conditions on g.

Assume that

(**CD**) the function $(x,y,u)\to g(t,x,y,u)$ is continuously differentiable for almost every $t\in[a,b]$.

The assumption (CD) is stronger than (L*), i.e., if (CD) holds, so does (L*).

Additionally, we assume that the function g satisfies:

(**I1**) $\exists\, c, m_g$ such that for almost every $t\in[a,b]$ and all $(x,y,u)\in S^*_\varepsilon(t)$

$$\begin{aligned}\det g_y(t,x,y,u)g_y(t,x,y,u)^T &\geq c>0\\ \text{and}\quad \left|\left[g_y(t,x,y,u)g_y(t,x,y,u)^T\right]^{-1}g_y(t,x,y,u)\right| &\leq m_g\end{aligned}$$

Then, for any $\lambda\in\mathbb{R}^n$, we have

$$\partial^L_{x,y,u}\,\langle\lambda,g(t,x,y,u)\rangle=\nabla g_{x,y,u}(t,x,y,u)^T\lambda$$

If $(\alpha, \beta_1, \beta_2 - \xi) \in \partial^L_{x,y,u} \langle \lambda, g(t,x,y,u) \rangle$ and $\xi \in N^C_U(u)$ then

$$\begin{aligned} \alpha &= g_x(t,x,y,u)^T \lambda \\ \beta_1 &= g_y(t,x,y,u)^T \lambda \\ \beta_2 &= g_u(t,x,y,u)^T \lambda + \xi \end{aligned}$$

We now make use of the assumption (**I1**). We know that $[g_y(t,x,y,u) g_y(t,x,y,u)^T]^{-1}$ exists, thus

$$\begin{aligned} \lambda &= \underbrace{[g_y(t,x,y,u) g_y(t,x,y,u)^T]^{-1} g_y(t,x,y,u)}_{\leq m_g} \beta_1 \\ \Longrightarrow \quad & |\lambda| \leq m_g \, |\beta_1| \leq m_g \, |(\beta_1, \beta_2)| \end{aligned}$$

This means that our new assumptions (**CD**) and (**I1**) imply (**A1**). As a consequence, we sum up our results:

Corollary 8.3.3 (Smooth maximum principle for index 1 case)
Suppose that (x^*, y^*, u^*) is a $W^{1,1}$-local minimizer for (P), the basic assumptions are satisfied, ($\mathbf{L}^*$) applies to f, (**CD**), (**I1**) apply to g. Then $\exists p \in W^{1,1}([a,b]; \mathbb{R}^n)$ and a $\lambda_0 \geq 0$ satisfying
the *nontriviality condition*: $\|p\|_\infty + \lambda_0 > 0$;
the *Euler adjoint inclusion*: for a measurable function $\xi(t) \in N^C_U(u^*(t))$ it holds

$$(-\dot{p}(t), 0, \xi(t)) \in \partial^C_{x,y,u} \langle p(t), f(t, x^*(t), y^*(t), u^*(t)) \rangle - \nabla g_{x,y,u}(t, x^*(t), y^*(t), u^*(t))^T \lambda(t) \text{ a.e.;}$$

the global *Weierstrass condition*: for almost every $t \in [a,b]$ and all $(x^*, y, u) \in S(t)$

$$\langle p(t), f(t, x^*(t), y, u) \rangle \leq \langle p(t), f(t, x^*(t), y^*(t), u^*(t)) \rangle \text{ a.e.;}$$

the *transversality condition*:

$$(p(a), -p(b)) \in N^L_E(x^*(a), x^*(b)) + \lambda_0 \partial^L l(x^*(a), x^*(b)).$$

Remark 8.3.4 (i) Observe that we do not impose smoothness assumptions on $(x,y,u) \mapsto f(x,y,u)$.

(ii) Under the assumptions of Proposition 8.3.3 there exists a $k_g > 0$ such that

$$|\lambda(t)| \leq m_g k_g L_f \, |p(t)| \text{ a.e.}$$

(iii) If $U = \mathbb{R}^k$, then the corollary holds with $N^C_U(u^*(t)) = \{0\}$.

(iv) A similar result was obtained in [23]. However, convexity was assumed while it is not necessary in our case.

8.4 An alternative "hybrid" result

So far the pair (y, u) has been treated as a control while only u was possibly subject to set control constraints. This fact may lead to different assumptions on y and u. We re-define the set control constraint $S(t)$ in (8.1) (and by doing so, also (8.2))

$$S(t,u) := \{(x,y,u) \ : \ g(t,x,y,u) = 0\};$$
$$S^*_\varepsilon(t,u) := \{(x,y,u) \in S(t,u) \ : \ |x(t) - x^*(t)| \leq \varepsilon\}$$

and the assumptions $(\mathbf{L}^*)$, **(A1)** into

$(\mathbf{L}^*_\mathbf{2})$ $\exists\, k_\psi > 0$ such that for almost every $t \in [a,b]$ and every $u \in U$, every (x_i, y_i) with $|x_i(t) - x^*(t)| \leq \varepsilon$:

$$|\psi(t,x_1,y_1,u) - \psi(t,x_2,y_2,u)| \leq k_\psi \left(|x_1 - x_2| + |y_1 - y_2|\right);$$

(A2) $\exists\, M > 0$ such that for (almost) every $t \in [a,b]$, all $(x,y,u) \in S^*_\varepsilon(t)$, all $\lambda \in \mathbb{R}^m$ and all $\xi \in N^L_U(u)$:

$$(\alpha,\beta) \in \partial^L_{x,y} \langle \lambda, g(t,x,y,u)\rangle \quad \Longrightarrow \quad |\lambda| \leq M\,|\beta|\,.$$

Remark 8.4.1 (i) The alternative assumption (L^*_2) requires f, g to be Lipschitz continuous in y but no Lipschitz continuity is imposed with respect to u.

(ii) Similarly to the differential case under the assumptions $(\mathbf{L}^*)$ and **(A1)**, we may utilize the differentiability of g. Suppose that the assumption

(CD2) The function $(x,y,u) \to g(t,x,y,u)$ is continuously differentiable for a.e. $t \in [a,b]$ and further, $(\alpha,\beta) \in N^C_{S(t,u)}(x,y)$.

Then there exists $\lambda : [a,b] \to \mathbb{R}^m$ such that

$$(\alpha,\beta) = g_{x,y}(t,x,y,u)^T \lambda(t).$$

This makes possible to state the following necessary conditions:

Corollary 8.4.2 Suppose that (x^*, y^*, u^*) is a $W^{1,1}$-local minimizer for (P), the basic assumptions are satisfied, (L_2^*) applies to f and g, moreover, (**CD2**) and (**A2**) apply to g then
$\exists\, p \in W^{1,1}([a,b];\mathbb{R}^n), \lambda_0 \geq 0$ satisfying the *Euler adjoint inclusion*

$$(-\dot{p}(t), 0) \in \partial^C_{x,y} \langle p(t), f(t, x^*(t), y^*(t), u^*(t))\rangle - g_{x,y}(t, x^*(t), y^*(t), u^*(t))^T \lambda(t) \quad \text{a.e.},$$

the *Global Weierstrass condition*: for almost every $t \in [a,b]$, all $u \in U$ and $(x^*(t), y) \in S(t,u)$

$$\langle p(t), f(t, x^*(t), y, u)\rangle \leq \langle p(t), f(t, x^*(t), y^*(t), u^*(t))\rangle\,.$$

8.5 Conclusion

One may ask whether we, in the smooth case, need both results of Corrolary 8.3.3 and the alternative Corrolary 8.4.2. Yes, in fact, both have their justification. The alternative version suits a broader class of problems. The first version imposes stricter assumptions, however, delivers potentially stronger results. This might be interesting for applications.

We again remark here that similar results can be obtained for higher index problems. This is the case when the algebraic equation reduces to $g(t, x(t)) = 0$. Our approach can nevertheless cover some of the higher index problems supposing that they satisfy $(A1)$.

Again we would like to emphasize the novelty of the presented approach such as the Implicit Functions theorem, customarily deployed with DAE problems, here did not need to be invoked in the first place.

Chapter 9

Constrained Control Problems with Differential Inclusions

9.1 Introduction

We have seen multifunctions appear in Chapters 6 and 8, somewhat tacitly, when we treat sets dependent on the time variable t: for example, the subdifferential or the normal cone to the graph of a function. Let us now make this concept more precise. A mapping $\Gamma : \mathbb{R}^m \to \mathcal{P}(\mathbb{R}^n)$ is called a *multifunction* (also a *set-valued function*) if it maps each $y \in \mathbb{R}^m$ to a subset $\Gamma(y)$ of $\mathbb{R}^n$ which may be also an empty set. We say that the multifunction Γ is *closed* (*compact*, *convex* or *nonempty*) on a set $S \subset \mathbb{R}^m$ if $\Gamma(y)$, the *image* of Γ, is closed (respectively, compact, convex or nonempty) for all $y \in S$.

We say that a multifunction $\Gamma : \mathbb{R}^m \to \mathcal{P}(\mathbb{R}^n)$ is *measurable* if the set

$$\{x \in \Omega : \Gamma(x) \cap C \neq \emptyset\} \tag{9.1}$$

is $\mathcal{L}$-measurable for every open set $C \subset \mathbb{R}^n$. The measurability of Γ can be defined equivalently if the set C in (9.1) is an arbitrary closed set.

We refer to literature (e.g. [13, 75]) for answers how the measurability of a multifunction is preserved under composition and limit-taking.

Another important concept, already used in Chapters 6 and 8, is the question whether the measurable and closed multifunction $\Gamma(y)$ has a *measurable selection* $\gamma(y)$, i.e. there exists an $\mathcal{L}$-measurable function $\gamma : S \to \mathbb{R}^n$ such that

$$\gamma(y) \in \Gamma(y) \quad \text{a.e. } y \in S.$$

Assume now that $S = [a,b]\times\mathbb{R}^n$, $F : [a,b]\times\mathbb{R}^n \to \mathbb{R}^n$ is a closed-valued multifunction, $x : [a,b] \to \mathbb{R}^n$ is an absolutely continuous function. Then a *differential inclusion* is defined by

$$\dot{x}(t) \in F(t,x(t)) \quad \text{a.e. } t \in [a,b]. \tag{9.2}$$

An absolutely continuous function x is a *solution* to (9.2) if there exists a measurable selection $\gamma(t)$ of $F(t,x(t))$ such that

$$x(t) = x(a) + \int_a^t \gamma(t)\, ds.$$

The differential inclusion (9.2) defines in a convenient way a dynamic system; the multifunction F describes the set of its possible velocities. Recall a Mayer type optimal control problem

$$(P_M) \quad \begin{cases} \text{Minimize } l(x(a),x(b)) \\ \text{subject to} \\ \quad \dot{x}(t) = f(t,x(t),u(t)) \quad \text{a.e. } t \in [a,b] \\ \quad u(t) \in U(t) \quad \text{a.e. } t \in [a,b] \\ \quad (x(a),x(b)) \in E. \end{cases}$$

A natural question is whether there exists an equivalent control problem with differential inclusions,

$$(P_{DI}) \quad \begin{cases} \text{Minimize } l(x(a),x(b)) \\ \text{subject to} \\ \quad \dot{x}(t) \in F(t,x(t)) \quad \text{a.e. } t \in [a,b] \\ \quad (x(a),x(b)) \in E, \end{cases}$$

where $F(t,x) := \{f(t,x,u) : u \in U(t)\}$. The positive answer to this question is a well-established result which makes use of the Generalized Filippov Selection Theorem (see e.g. [13, 75]).

We turn our attention to the extended problem case when (P_M) additionally contains a mixed state constraint $g(t,x(t),u(t)) \le 0$ for almost every $t \in [a,b]$. In what follows in this chapter we investigate the existence of a solution and the necessary conditions for optimality for a control problem described by a differential inclusion

$$\dot{x}(t) \in F^-(t,x(t)) \quad \text{a.e. } t \in [a,b]$$

with a suitable multifunction F^- such that the constraint $g(t,x(t),u(t)) \le 0$ is already incorporated in the definition of F^-.

In this respect, issues on measurability, convexity, compactness of trajectories and Lipschitz continuity of F^- need to be addressed. Quite handily, even if a given optimal control problem is not

formulated as a differential inclusion problem but rather in terms of "conventional", functional relations, it is possible to make a statement on the existence of a minimizer for the original "conventional" problem due to the results obtained for the diffential inclusion problems.

Most of the the results presented here may have appeared earlier in the literature. See e.g. [23], Chapter 2 of [75] and the references therein. However, the novelty lies in presenting these results in a new, concise and clarifying way.

9.2 Auxiliary definitions

Suppose we have a control system with mixed constraints as inequalities,

$$(C) \quad \begin{cases} \dot{x}(t) = f(t, x(t), u(t)), & \text{a.e. } t \in [a, b], \\ g(t, x(t), u(t)) \leq 0, & \text{a.e. } t \in [a, b], \\ u(t) \in U, & \text{a.e. } t \in [a, b], \\ (x(a), x(b)) \in E, \end{cases} \tag{9.3}$$

where both functions $f : [a, b] \times \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^n$ and $g : [a, b] \times \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^m$ as $\mathcal{L} \times \mathcal{B} \times \mathcal{B}$-measurable, the set $U \subset \mathbb{R}^k$ is closed and E is a closed subset of $\mathbb{R}^n \times \mathbb{R}^n$. The function $x : \mathbb{R} \to \mathbb{R}^n$ is absolutely continuous and $u : \mathbb{R} \to \mathbb{R}^k$ is measurable.

Suppose that $S : [a, b] \to \mathbb{R}^n \times \mathbb{R}^k$ is a closed-valued multifunction. Define the set S as

$$S(t) := \{(x, u) \in \mathbb{R}^n \times U \ : \ g(t, x, u) \leq 0\}.$$

Then the expression

$$(x(t), u(t)) \in S(t), \quad \text{a.e. } t \in [a, b]$$

is a mixed state constraint in general form, equivalent to $g(t, x(t), u(t)) \leq 0$, a.e. $t \in [a, b]$ (see also (2.9)). For a given (t, x), the multifunction S^- describes the set of controls u satisfying the control set constraint,

$$S^-(t, x) := \{u \in U \ : \ (x, u) \in S(t)\}.$$

Note that for each $t \in [a, b]$, the set $S(t)$ is the graph of the multifunction $x \to S^-(t, x)$ and

$$u \in S^-(t, x) \quad \Longleftrightarrow \quad (x, u) \in S(t).$$

Now we define the multifunctions $F : [a,b] \times \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^m$ and $F^- : [a,b] \times \mathbb{R}^n \to \mathbb{R}^n$,

$$F(t,x) := \{(f(t,x,u), g(t,x,u)) \ : \ u \in U\},$$
$$F^-(t,x) := \left\{f(t,x,u) \ : \ u \in S^-(t,x)\right\}.$$

The control system (C) can now be expressed in terms of differential inclusions:

$$\begin{cases} \dot{x}(t) \in F^-(t,x(t)) \quad \text{a.a. } t \in [a,b] \\ (x(a),x(b)) \in E \end{cases} \tag{DI}$$

Let the multifunction X denote a *tube* around a feasible trajectory x^* at any time $t \in [a,b]$,

$$X(t) := x^*(t) + \varepsilon\overline{B}.$$

Then a set

$$S^*_\varepsilon(t) := \left\{(x,u) \in \mathbb{R}^n \times \mathbb{R}^k \ : \ (x,u) \in S(t)\right\} \bigcap (X(t) \times U)$$

describes all feasible processes which are close enough to a particular *reference feasible process* (x^*,u^*). Now and later we will call $x \in W^{1,1}([a,b];\mathbb{R}^n)$ an F^- *feasible trajectory*, iff

$$x(t) \in X(t), \quad \dot{x}(t) \in F^-(t,x(t))$$

at almost every $t \in [a,b]$.

9.3 Main assumptions

Let G be the closed graph of a multifunction $\Gamma : \mathbb{R}^n \to \mathbb{R}^k$ with $(x^*,u^*) \in G$. The following condition imposed on G is known as the *bounded slope condition* (see [13] for reference):

(BS') There exist $\varepsilon > 0$, $R' > 0$ and measurable function $\mathcal{K}$ such that, for almost all t, all $x \in B(x^*(t),\varepsilon)$ and all $u \in B(u^*(t),R')$ we have

$$(\alpha,\beta) \in N^P_G(x,u) \quad \Longrightarrow \quad |\alpha| \le \mathcal{K}(t)\,|\beta|\,.$$

The following theorem will assert that if a multifunction satisfies (BS') it is also pseudo-Lipschitz:

Theorem 9.3.1 (Thm. 3.5.2 in [15]) Let Γ satisfy (BS') near the point $(x_0,v_0) \in G = Gr\,\Gamma$.

Then for any $\xi \in (0,1)$ and any $x_1, x_2 \in B(x_0, \varepsilon_\xi)$ the following holds

$$\Gamma(x_1) \cap \overline{B}(u_0, (1-\xi)R') \subset \Gamma(x_2) + k\,|x_2 - x_1|\,\overline{B},$$

where $\varepsilon_\xi = \min\{\varepsilon, \xi R'/(3k)\}$.

Let (x^*, u^*) be a reference feasible process for the control system (C). The assumptions we operate with are:

(A1) (The *Lipschitz* properties of f and g)

There exist integrable functions $k_x^f, k_u^f : [a,b] \to \mathbb{R}$ (and k_x^g, k_u^g, respectively) such that, for every two state-control pairs $(x_1, u_1), (x_2, u_2)$ and almost every $t \in [a,b]$ the functions f, g (jointly denoted as ϕ) satisfy the following condition

$$|\phi(t, x_1, u_1) - \phi(t, x_2, u_2)| \le k_x^\phi(t)\,|x_1 - x_2| + k_u^\phi(t)\,|u_1 - u_2|\,,$$

for all x_i subject to $|x_i - x^*(t)| \le \varepsilon$, $u_i \in U$.

(A2) The control set $U \subset \mathbb{R}^k$ is closed. For all $u \in U$, $|u| \le R_u$ with some $R_u > 0$.

(A3) (The *bounded slope* condition) There exists an integrable function $M : [a,b] \to \mathbb{R}_+$ such that for all $(x,u) \in S_\varepsilon^*(t)$, $\eta \in N_U^L(u)$, $\gamma \in \mathbb{R}_+^m$, $\langle \gamma, g(t,x,u)\rangle = 0$ we have, at almost every $t \in [a,b]$,

$$(\alpha, \beta - \eta) \in \partial_{x,u}^L \langle \gamma, g(t,x,u)\rangle \quad \Longrightarrow \quad |\gamma| \le M(t)\,|\beta|\,.$$

(NE) (*Nonemptyness*) For each $t \in [a,b]$ and $x \in \mathbb{R}^n$, there exists $u \in U$ such that $g(t,x,u) \le 0$.

(C) (*Convexity*) For all $(t,x) \in [a,b] \times X(t)$, the set $F^-(t,x)$ is convex.

Remark 9.3.2 For almost every $t \in [a,b]$ and every $u \in U$, we obtain from (A1) the estimate

$$|f(t, x^*(t), u) - f(t, x^*(t), u^*(t))| \le k_u^f(t)\,|u - u^*(t)|\,, \tag{9.4}$$

and further, via the triangle inequality,

$$|f(t, x^*(t), u)| - |f(t, x^*(t), u^*(t))| \le k_u^f(t)\,|u - u^*(t)|\,,$$

leading to $|f(t,x^*(t),u)| \le k_u^f(t)\,|u-u^*(t)| + |\dot{x}^*(t)|$. Since U is compact, there exists $R_u > 0$ such that $|u| < R_u/2$, $\forall u \in U$. Thus $|f(t,x^*(t),u)|$ is bounded by an integrable $k(t)$,

$$|f(t,x^*(t),u)| \le \underbrace{k_u^f(t)R_u + |\dot{x}^*(t)|}_{:=k(t)}. \tag{9.5}$$

We also conclude that the sets

$$f(t,x,U) := \{f(t,x,u) \,:\, u \in U\} \qquad \text{and} \qquad g(t,x,U) := \{g(t,x,u) \,:\, u \in U\}$$

are compact for all $x \in x^*(t) + \varepsilon B$ and $t \in [a,b]$, since U is compact and f, g are continuous.

9.4 On the convexity of $F^-(t,x)$

The assumed convexity of $F^-(t,x)$ may be many times difficult to verify. Therefore we like to know whether the convexity of $F^-(t,x)$ can be checked indirectly and therefore are looking for helpful connections between F^- and the other multifunctions. Let the multifunctions F^f, F^g be defined as follows:

$$F^f(t,x,U) := \{f(t,x,u) \,:\, u \in U\},$$
$$F^g(t,x,U) := \{g(t,x,u) \,:\, u \in U\},$$

for all $t \in [a,b], x \in \mathbb{R}^n$.

Then the convexity dependencies of F, F^-, S^-, F^f, F^g are shown in the following diagram.

$$\begin{array}{lcl}
F(t,x) \text{ convex} & \begin{array}{c}\Longrightarrow\\ \not\Longleftarrow\end{array} & F^f(t,x),\ G^g(t,x) \text{ convex} \\
\Downarrow \ \Uparrow & & \Uparrow \ \Downarrow \\
F^-(t,x) \text{ convex} & \Longleftrightarrow & F^-(t,x) \text{ convex} \quad \begin{array}{c}\not\Longrightarrow\\ \not\Longleftarrow\end{array} \quad S^-(t,x) \text{ convex}
\end{array}$$

We make the exercise of proving these relationships.

(a) F **convex** $\Longrightarrow$ F^- **convex**

Suppose $\gamma_1, \gamma_2 \in F(t,x)$, i.e. there are controls $u_1, u_2 \in U$ with

$$\begin{aligned}\gamma_1 &= [f(t,x,u_1), g(t,x,u_1)],\\ \gamma_2 &= [f(t,x,u_2), g(t,x,u_2)]\end{aligned}$$

Assume now that F is convex, i.e. $\forall \gamma_1, \gamma_2 \in F(t,x)$, $\forall \alpha \in [0,1]$: $\alpha\gamma_1 + (1-\alpha)\gamma_2 \in F(t,x)$ or, equivalently,

$$\exists \tilde{u}: \quad \begin{bmatrix} \alpha f(t,x,u_1) + (1-\alpha) f(t,x,u_2) \\ \alpha g(t,x,u_1) + (1-\alpha) g(t,x,u_2) \end{bmatrix}^T = \begin{bmatrix} f(t,x,\tilde{u}) \\ g(t,x,\tilde{u}) \end{bmatrix}^T. \tag{9.6}$$

Suppose, on the other hand, $\gamma_1, \gamma_2 \in F^-(t,x)$, that is, $\exists\, u_1, u_2 \in U$

$$\begin{aligned}\gamma_1 &= \{f(t,x,u_1) \,:\, g(t,x,u_1) \le 0\},\\ \gamma_2 &= \{f(t,x,u_2) \,:\, g(t,x,u_2) \le 0\}.\end{aligned}$$

Then for any $\beta \in [0,1]$ and the functions $(t,x) \mapsto g(t,x,u_1)$ and $(t,x) \mapsto g(t,x,u_2)$ we have

$$\beta g(t,x,u_1) + \underbrace{(1-\beta)}_{\ge 0} g(t,x,u_2) \le 0.$$

From (9.6) we know that $\exists\, \tilde{u}$ such that $\beta g(t,x,u_1) + (1-\beta) g(t,x,u_2) = g(t,x,\tilde{u})$ and with the same $\tilde{u}$ there is a $\gamma \in F^-(t,x)$,

$$\begin{aligned}\gamma &= \{f(t,x,\tilde{u}) \,:\, g(t,x,\tilde{u}) \le 0\}\\ &= \{\beta f(t,x,u_1) + (1-\beta) f(t,x,u_2) \,:\, \beta g(t,x,u_1) + (1-\beta) g(t,x,u_2) \le 0\}\end{aligned} \tag{9.8}$$

It follows that for any $\beta \in [0,1]$ and any $\gamma_1, \gamma_2 \in F^-(t,x)$

$$\beta\gamma_1 + (1-\beta)\gamma_2 \in F^-(t,x),$$

i.e. $F^-(t,x)$ is convex.

(b) Let a counterexample explain why the convexity of F^- does not imply the convexity of F. Let $f(t,x,u) = u(t)$, $g(t,x,u) = u^2(t) - 1$ and

$$F^-(t,x) := \left\{u \,:\, u^2 - 1 \le 0\right\}.$$

Obviously, $F^-(t,x)$ is equal to $[-1,1]$ and convex. However, for $F(t,x) = \{(u, u^2-1) \,:\, u \in \mathbb{R}\}$, its image is the parabola $u^2 - 1$ and the convexity does not hold.

(c) F **convex** $\Longrightarrow$ F^f, F^g **convex**
To see the implication, one needs once again to formalize convexity,

$$\forall u_1, u_2 \in U,\ \forall \alpha \in [0,1] : \exists\, \tilde{u} : \beta f(t,x,u_1) + (1-\beta) f(t,x,u_2) = f(t,x,\tilde{u}) \tag{9.9}$$
$$\forall v_1, v_2 \in U,\ \forall \beta \in [0,1] : \exists\, \tilde{v} : \beta g(t,x,v_1) + (1-\beta) g(t,x,v_2) = g(t,x,\tilde{v}) \tag{9.10}$$

and see that, due to (9.6), $\tilde{u}, \tilde{v}$ do exist and in case $u_1 = v_1, u_2 = v_2$ we have $\tilde{u} = \tilde{v}$.

(d) F^f, F^g **convex** $\not\Longrightarrow$ F **convex**
The counterexample is

$$f(t,x,u) = u^2,$$
$$g(t,x,u) = u$$

on the interval $[0,1]$. The resulting multifunction is

$$F(t,x) = \left\{(u^2, u) \ :\ u \in [0,1]\right\} \subset \mathbb{R}^2,$$

corresponds to $Gr\sqrt{x}$ plotted on the (x,y)-plane. It is nonconvex.

(e) F^f, F^g **convex** $\not\Longrightarrow$ F^- **convex**
Define

$$F^f(t,x,U) := \{u \ :\ u \in [0,4\pi]\} = [0,4\pi],$$
$$F^g(t,x,U) := \{\sin u \ :\ u \in [0,4\pi]\} = [-1,1].$$

Both F^f, F^g are convex. Nevertheless,

$$\begin{aligned} F^-(t,x) &= \{u \ :\ \sin u \le 0,\ u \in [0,4\pi]\} \\ &= \{u \ :\ u \in [\pi, 2\pi] \cup [3\pi, 4\pi]\} \\ &= [\pi, 2\pi] \cup [3\pi, 4\pi] \end{aligned}$$

is not convex.

(f) F^- **convex** $\not\Longrightarrow$ F^f, F^g **convex**
Define $U := [-1,1]$ and

$$f(t,x,u) := u,$$
$$g(t,x,u) := (-u, u^3 - u).$$

Then both $S^-(t,x) = [0,1]$ and $F^-(t,x) = [0,1]$ are convex. On the other hand, although $F^f(t,x) = [0,1]$ is convex, we do not have convexity of $F^g(t,x) = \{(-u, u^3 - u) : \ u \in [-1,1]\}$.

(g) F^- **convex** $\not\Longrightarrow$ S^- **convex**
Define

$$g(t,x,u) := 1 - u^2, \quad u \in [-2,2].$$

Therefore

$$\begin{aligned} S^-(t,x) :&= \{u \in [-2,2] \ : \ g(t,x,u) \leq 0\} \\ &= \left\{u \in [-2,2] \ : \ -u^2 + 1 \leq 0\right\} \\ &= [-2,-1] \cup [1,2]. \end{aligned}$$

and

$$F^-(t,x) := \left\{|u| \ : \ u \in S^-(t,x)\right\} = [1,2].$$

So $F^-(t,x)$ is convex while $S^-(t,x)$ is not.

(h) S^- **convex** $\not\Longrightarrow$ F^- **convex**
Define $S^-(t,x) = [-1,0]$. It is convex but $F^-(t,x) = \{(u,|u|) \ : \ u \in S^-(t,x)\}$ is not.

9.5 Main results

Lemma 9.5.1 Assume that (A1) is satisfied for f and g and (A2), (NE), (C) hold as well. Then, for each $t \in [a,b]$, the following holds:

(i) The set $S^*_\varepsilon(t)$ is nonempty and compact.

(ii) For every $x \in \mathbb{R}^n$, the sets $S^-(t,x)$ and $F^-(t,x)$ are nonempty and compact.

(iii) The multifunction $t \to S(t)$ is measurable.

Proof: (i) For every $t \in [a,b]$, $S^*_\varepsilon(t)$ is nonempty due to the assumption (NE). Let $\{(x_i,u_i)\} \subset S^*_\varepsilon(t)$ be a convergent sequence, $(x_i,u_i) \to (x,u)$. Then $g(t,x_i,u_i) \leq 0$ for all $i \in \mathbb{N}$ and, since g is continuous,

$$g(t,x_i,u_i) \xrightarrow{(x_i,u_i)\to(x,u)} g(t,x,u).$$

Obviously, $g(t,x,u) \leq 0$. Indeed, suppose $g(t,x,u) > 0$, then $\exists\, n \in \mathbb{N}$ such that $g(t,x_m,u_m) > 0$ for all $m \geq n$ which contradicts the assumption. Hence, $(x,u) \in S^*_\varepsilon(t)$ and $S^*_\varepsilon(t)$ is closed. $X(t) \times U$ is bounded, so $S^*_\varepsilon(t)$ is compact.

(ii) The nonemptyness of $S^-(t,x)$ and $F^-(t,x)$ follows directly from the assumption (NE). Note that $S^-(t,x)$ is closed since it can be viewed as $g^{-1}\left(\{g(t,x,u) \,:\, u \in U\} \cap \mathbb{R}^m_{\leq 0}\right)$: The image of $u \mapsto g(t,x,u)$ is compact since U is compact and g is continuous. Its intersection with the closed halfspace $\mathbb{R}^m_{\leq 0}$ is also a closed set. The inverse image of a closed set for a continuous function is, again, a closed set. On the other hand, $S^-(t,x)$ is bounded, since U is bounded. Hence, $S^-(t,x)$ is compact.

Compactness of $F^-(t,x)$ follows from the compactness of $S^-(t,x)$ since $u \mapsto f(t,x,u)$ is continuous.

(iii) The multifunction $t \to S(t)$ is a level-set mapping. Proposition 14.33 in [70] ensures that $t \to S(t)$ is measurable as long as the function $t \mapsto g(t,x,u)$ is measurable which is exactly our case. ∎

Remark 9.5.2 Under the previous assumption of $S(t)$ being a closed-valued set, (iii) of the last lemma is equivalent to the graph of S being a $\mathcal{L} \times \mathcal{B}$ set. This is a direct application of Theorem 2.3.7 in [75].

Lemma 9.5.3 Assume (A1), (A2), (NE), (C) as in Lemma 9.5.1 and that (9.5) holds. The multifunction F^- has the following properties:

(i) $(t,x) \to F^-(t,x)$ is $\mathcal{L} \times \mathcal{B}$ measurable.

(ii) For almost all $t \in [a,b]$ and every $x \in X(t)$, there exists an integrable function c such that for all $\gamma \in F^-(t,x)$, $|\gamma| \leq c(t)$.

(iii) The graph of $x \to F^-(t,x)$, when restricted to $X(t)$, is a closed set.

PROOF: (i) By assumption, $t \to f(t,x,u)$ is $\mathcal{L}$ measurable and $(x,u) \to f(t,x,u)$ is continuous, thus $\mathcal{B}$ measurable. Select an arbitrary open set $A \in \mathrm{dom} f$, i.e. $f(A) \neq \emptyset$. Since f is $\mathcal{L} \times \mathcal{B}$ measurable, the inverse image $f^{-1}(A)$ is a $\mathcal{L} \times \mathcal{B}$ measurable set. Since $t \to S(t)$ is closed-valued and measurable by Lemma 9.5.1, its graph

$$\Sigma := \{(t,x,u) \in [a,b] \times \mathbb{R}^n \times U \,:\, (x,u) \in S(t)\}\,,$$

is, by Remark 9.5.2, a $\mathcal{L} \times \mathcal{B}$ measurable set. Since countable intersections of measurable sets are measurable, $f^{-1}(A) \cap \Sigma$ is a measurable set. Considering again Remark 9.5.2 and the fact that $f^{-1}(A) \cap \Sigma$ is the graph of

$$F^-(A) = \left\{(t,x) \in [a,b] \times \mathbb{R}^n \ : \ \exists u \in U : \ (t,x,u) \in f^{-1}(A) \cap \Sigma\right\},$$

yields the measurability of F^-.

(ii) Remark 9.3.2 established that, for all $(x,u) \in X(t) \times U$, there is an integrable function k such that $|f(t,x,u)| \leq k(t)$. Immeadiately, it follows for all γ with

$$\gamma \in \left\{f(t,x,u) \ : \ (x,u) \in X(t) \times S^-(t,x)\right\},$$

at almost all $t \in [a,b]$, that $|\gamma| \leq c(t)$ where c is an integrable function less or equal to k.

(iii) For almost all $t \in [a,b]$, we have

$$Gr\left[x \to F^-(t,x)\right] \cap X(t) = \left\{(x, f(t,x,u)) \ : \ (x,u) \in X(t) \times S^-(t,x)\right\}.$$

The claim is clear since $X(t) \times S^-(t,x)$ is a compact set. ∎

Now we are ready to formulate a *Compactness of Trajectories Theorem* which describes the closure properties of the set of F^- trajectories. The following theorem is a specific case of Theorem 2.5.3 in [75].

Theorem 9.5.4 Assume, as in Lemma 9.5.1, (A1), (A2), (NE) and (C). Take any sequence $\{x_i\}$ of F^- trajectories, i.e.

$$x_i \in W^{1,1}([a,b];\mathbb{R}^n), \quad x_i(t) \in X(t), \quad \dot{x}_i(t) \in F^-(t,x_i(t)) \text{ at a.e. } t \in [a,b]$$

such that $x_i(a) \in X(a)$. Then there is a subsequence (without relabeling) such that

$$x_i \to x \text{ uniformly} \quad \text{and} \quad \dot{x}_i \to \dot{x} \text{ weakly in } L^1$$

for some $x \in W^{1,1}([a,b];\mathbb{R}^n)$ such that

$$\dot{x}(t) \in F^-(t,x(t)) \text{ at a.e. } t \in [a,b],$$

i.e. the limit of $\{x_i\}$ is itself an F^- feasible trajectory.

We now have all ingredients to obtain the relation between the set of F^- feasible trajectories and the set of feasible trajectories for the problem (9.3). Let $\mathcal{S}^*_{[a,b]}$ denote the set of all absolutely continuous functions $x \in X(t)$ such that x and a control function $u : [a,b] \to U$ are an admissible process for (9.3). On the other hand, we define the set of all F^- *feasible trajectories associated with* E by

$$\mathcal{R}^*_{[a,b]}(E) := \left\{ x \in W^{1,1}([a,b];\mathbb{R}^n) \ : \ x \text{ is an } F^- \text{ trajectory and } (x(a),x(b)) \in E \right\}.$$

Theorem 9.5.5 Assume, as in Lemma 9.5.1, (A1), (A2), (NE), (C) and that $E \subset \mathbb{R}^n \times \mathbb{R}^n$ is a closed set. Then $x \in \mathcal{S}^*_{[a,b]}(E)$ if and only if $x \in \mathcal{R}^*_{[a,b]}(E)$.

Proof: The implication

$$x \in \mathcal{S}^*_{[a,b]}(E) \quad \Longrightarrow \quad x \in \mathcal{R}^*_{[a,b]}(E)$$

is trivial. The opposite direction is validated by the Filippov Selection Theorem. The precise arumentation can be found in [26, 45]. ∎

We will now focus on the Lipschitz properties of F^- trajectories and the relevance of the previously made assumptions herein. As we see in the following example, (A1), (A2) and (NE) without (A3) do not asset the lower semicontinuity of $x \to S^-(t,x)$.

Example 9.5.6 Suppose that $g(x,u) = |x|\,u$, $x,u \in \mathbb{R}$, $u \in [-1,1]$. Suppose further that$x \in X_\varepsilon := [-\varepsilon,\varepsilon]$ for some $\varepsilon > 0$. We verify that the assumptions (A1), (A2) are satisfied:

(A1):

$$\begin{aligned}
|g(t,x_1,u_1) - g(x_2,u_2)| &= ||x_1|\,u_1 - |x_2|\,u_2| \\
&= ||x_1|\,u_1 - |x_1|\,u_2 + |x_1|\,u_2 - |x_2|\,u_2| \\
&= ||x_1|\,(u_1-u_2) + u_2(|x_1| - |x_2|)| \\
&\le ||x_1|\,(u_1-u_2)| + |u_2(|x_1| - |x_2|)| \\
&\le k_u\,|u_1-u_2| + k_x\,|x_1-x_2|\,, \quad \text{with } k_x := \max|u_2|\,, k_u := \max|x_1|
\end{aligned}$$

(A2): The set $[-1,1]$ is compact.

Thus, $S(t,x) = \{u \in [-1,1] \ : \ |x|\,u \le 0\}$ and

$$S^-(t,x) = \begin{cases} [-1,1], & \text{if } x = 0, \\ [-1,0]\,, & \text{if } x \neq 0 \end{cases}$$

This multifunction is, clearly, not l.s.c. in 0: select an open set $\mathcal{U} = (0,1) \subset S^-(t,0)$. For all $\eta > 0$ there is no x' in $(-\eta,\eta)$ such that $S^-(t,x') \cap \mathcal{U} \neq \emptyset$.

Assumption (A3) excludes this example from our context. Consider (A3) with $x = 0$, $u = 1/2$, $\gamma = 1$ and $(\alpha, \beta) = (1/3, 0)$. Then $N^L_{[-1,1]}(1/2) = \{0\}$ and we observe that

$$(\tfrac{1}{3}, 0) \in \partial^L_{(x,u)} \left\langle 1, g(0, \tfrac{1}{2}) \right\rangle = (u[-1, 1], |x|)\Big|_{x=0,u=1/2} = ([-\tfrac{1}{2}, \tfrac{1}{2}], 0).$$

However, for any $M > 0$ we have $1 > M \cdot 0 = 0$. So (A3) does not hold.

Lemma 9.5.7 (A3) implies (BS').

Proof: Our procedure is to prove that (A3) implies (BS), a bounded slope condition on $t \to S^*_\varepsilon(t)$ as defined below. (BS), in its turn, will imply (BS').

(BS) There exists a measurabe function $\mathcal{K} : [a, b] \to \mathbb{R}_+$ such that, for almost all $t \in [a, b]$ and all $(x, u) \in S^*_\varepsilon(t)$ the following implication holds

$$(\alpha, \beta) \in N^P_{S^*_\varepsilon(t)}(x, u) \quad \Longrightarrow \quad |\alpha| \le \mathcal{K}(t)\, |\beta| \,,$$

The following result is a helpful first step in that procedure.

Lemma 9.5.8 (Characterization of $N^L_{S^*_\varepsilon(t)}$) For almost every $t \in [a, b]$, for all $(x, u) \in S^*_\varepsilon(t)$ and all $(\alpha, \beta) \in N^L_{S^*_\varepsilon(t)}(x, u)$, there exists an $r \ge 0$ with $\langle r, g(t, x, u)\rangle = 0$ such that

$$(\alpha, \beta) \in \partial^L_{(x,u)} \langle r, g(t, x, u)\rangle + \{0\} \times N^L_U(u).$$

Proof: We introduce some naming conventions. For a fixed $t \in [a, b]$ consider the function

$$G_t(x, u) := g(t, x, u)$$

and define $D := \{\xi \in \mathbb{R}^m \,:\, \xi \le 0\}$, the halfspace of $\mathbb{R}^m$. Furthermore, define

$$C_1(t) := G_t^{-1}(D) = \left\{(x, u) \in \mathbb{R}^n \times \mathbb{R}^k \,:\, g(t, x, u) \le 0\right\},$$

and

$$C_2(t) := X(t) \times U.$$

This allows us to write $S^*_\varepsilon(t) = C_1(t) \cap C_2(t)$.

Select any $(x, u) \in S^*_\varepsilon(t)$ and any $r \in N^L_D(G_t(x, u))$. The proximal normal inequality for D, taking into account that D is closed and convex (thus, $N^L_D(x, u) = N^P_D(x, u)$), and G_t is Lipschitz continuous,

holds as:

$$\langle r, \xi - G_t(x,u)\rangle \leq 0.$$

for all $\xi \in D$. Select the set of elements $\{\xi^1, \ldots, \xi^m\} \subset D$ such that ξ_i^j, the i-th component of ξ^j equals

$$\xi_i^j = \begin{cases} 0, & i = j, \\ G_{t,i}(x,u), & i \neq j, \end{cases}$$

where $G_{t,i}(x,u)$ is the i-th component of $G_t(x,u)$. Thus,

$$\xi^j - G_t(x,u) = \begin{cases} 0, & \text{in every component } i \neq j, \\ G_{t,j}(x,u), & \text{in the } j\text{-th component} \end{cases}, \quad \text{for all } j = 1, \ldots, m.$$

Then plugging $\{\xi^1, \ldots, \xi^m\}$ into the proximal normal inequality for D gives us

$$r_j \cdot G_{t,j}(x,u) \leq 0, \quad \text{for all } j = 1, \ldots, m \quad \Longleftrightarrow \quad r \geq 0.$$

With $r \geq 0$ and $\xi = 0$ we verify that $\langle r, G_t(x,u)\rangle = 0$.

In the next step, we conclude, by contradiction, that the only $r \in N_D^L(G_t(x,u))$ satisfying

$$0 \in \partial^L_{(x,u)} \langle r, g(t,x,u)\rangle$$

is $r = 0$: Suppose there exists $r > 0$ with $\langle r, G_t(x,u)\rangle = 0$. Since we always have $\eta = 0 \in N_U^L(u)$, we obtain from (A3) with setting $(\alpha, \beta) = (0,0)$:

$$(0,0) \in \partial^L_{(x,u)} \langle r, g(t,x,u)\rangle \quad \Longrightarrow |r| \leq M(t) \cdot 0,$$

a result contradicting the assumption in (A3).

We have shown that $r = 0$. This fact and and $C_1(t)$ being closed and strictly continuous allows us to apply Corollary 10.50 in [70] which concludes the inclusion

$$N^L_{C_1(t)}(x,u) \subset \bigcup \left\{ \partial^L_{(x,u)} \langle r, g(t,x,u)\rangle \ : \ r \in N_D^L(G_t(x,u)) \right\}.$$

This means there exists some $r \in N_D^L(G_t(x,u))$ with, as previously shown $r \geq 0$ and $\langle r, g(t,x,u)\rangle = 0$, such that

$$\text{if } v_1 \in N^L_{C_1(t)}(x,u), \quad \text{then } v_1 \in \partial^L_{(x,u)} \langle r, g(t,x,u)\rangle . \tag{9.11}$$

By the definition of $X(t)$ we can always make sure that $N^L_{X(t)}(x) = \{0\}$. Applying Corollary 2.4.5 in [13] for $C_2(t) = X(t) \times U$ we have

$$N^L_{C_2(t)}(x,u) = \{0\} \times N^L_U(u). \tag{9.12}$$

We now prove that $N^L_{C_1(t)}(x,u)$ and $N^L_{C_2(t)}(x,u)$ are *transversal*, i.e.,

$$-N^L_{C_1(t)}(x,u) \cap N^L_{C_2(t)}(x,u) = \{(0,0)\}.$$

Let

$$(x,u) \in C_1(t) \cap C_2(t) \quad \text{and} \quad (\alpha,\eta) \in -N^L_{C_1(t)}(x,u) \cap N^L_{C_2(t)}(x,u),$$

wherefrom, clearly, follows

$$(\alpha,\eta) \in -N^L_{C_1(t)}(x,u) \tag{9.13}$$

and

$$(\alpha,\eta) \in N^L_{C_2(t)}(x,u). \tag{9.14}$$

From (9.12) and (9.14) we obtain that $\alpha = 0$ and $\eta \in N^L_U(u)$. Together with (9.13) it implies that

$$(0,-\eta) \in N^L_{C_1(t)}(x,u).$$

It follows with (9.11) that there exists an $r \geq 0$ such that

$$(0,-\eta) \in \partial^L_{(x,u)} \langle r, g(t,x,u)\rangle$$

and $\langle r, g(t,x,u)\rangle = 0$. It allows to invoke the assumption (A3) in this setting, with $\beta = 0$, as follows:

$$(\alpha, \beta-\eta) = (0,-\eta) \in \partial^L_{(x,u)} \langle r, g(t,x,u)\rangle \quad \Rightarrow \quad |r| \leq M(t)\,|\beta| = 0.$$

Therefore, $r = 0$ and

$$\partial^L_{(x,u)} \langle 0, g(t,x,u)\rangle = (0,0).$$

This shows that η can only be 0. In other words,

$$(\alpha,\eta) = (0,0),$$

which proves the transversality of $N^L_{C_1(t)}$ and $N^L_{C_2(t)}$.

The property of the transversality allows us to apply Theorem 11.39 in [18] leading to

$$N^L_{S^*_\varepsilon(t)}(x,u) = N^L_{C_1}(x,u) \cap N^L_{C_2}(x,u) \subset N^L_{C_1(t)}(x,u) + N^L_{C_2(t)}(x,u).$$

From the last inclusion we deduce that for all $(\alpha,\beta) \in N^L_{S^*_\varepsilon(t)}(x,u)$,

$$(\alpha,\beta) \in N^L_{C_1(t)}(x,u) + \{0\} \times N^L_U(u),$$

and, subsequently, there exists $r \geq 0$ with $\langle r, g(t,x,u)\rangle = 0$ such that

$$(\alpha,\beta) \in \partial^L_{(x,u)} \langle r, g(t,x,u)\rangle + \{0\} \times N^L_U(u). \qquad \blacksquare$$

Proof (of Lemma 9.5.7, continued): Assume, as in (BS'), that $(\alpha,\beta) \in N^P_{S(t)}(x,u)$. Then it follows, for any $\gamma \geq 0$ with $\langle \gamma, g(t,x,u)\rangle = 0$, by the assumption (A3) and Lemma 9.5.8 that

$$\begin{aligned}
&(\alpha,\beta) \in \partial^L_{(x,u)} \langle \gamma, g(t,x,u)\rangle + \{0\} \times N^L_U(u) \\
\iff &(\alpha, \beta-\eta) \in \partial^L_{(x,u)} \langle \gamma, g(t,x,u)\rangle, \quad \eta \in N^L_U(u) \\
\implies &|\gamma| \leq M(t)\,|\beta|\,.
\end{aligned}$$

It can be shown that for all $(\alpha,\beta) \in N^P_{S(t)}(x,u)$, $\eta \in N^L_U(u)$ and γ as defined above, the following estimate holds:

$$|(\alpha, \beta-\eta)| \leq |\gamma|\,\mathcal{K}(t),$$

where $\mathcal{K}(t)$ is a measurable function.

Together with $|\gamma| \leq M(t)\,|\beta|$ it gives us

$$|\alpha| \leq |(\alpha,\beta-\eta)| \leq |\gamma|\,K^g(t) \leq K^g(t)M(t)\,|\beta|$$

and setting $\mathcal{K}(t) := K^g(t)M(t)$ we obtain $|\alpha| \leq \mathcal{K}(t)\,|\beta|$. It proves $(A3) \Rightarrow (BS)$.

Leaving out a number of technical steps here, one explores the property of normal cones to closed set,

$$N^P_{S(t)}(x,u) \subset N^P_{S^*_\varepsilon(t)}(x,u).$$

Hereby we achieve that, for all (x,u) and almost every $t \in [a,b]$, (BS) implies there exists a measurable function $\mathcal{K} : [a,b] \to \mathbb{R}_+$ such that

$$(\alpha,\beta) \in N^P_{S(t)}(x,u) \quad \Longrightarrow \quad (\alpha,\beta) \in N^P_{S^*_\varepsilon(t)}(x,u) \quad \Longrightarrow \quad |\alpha| \leq \mathcal{K}(t)\,|\beta|\,. \tag{9.15}$$

$\blacksquare$

In other words, if $(x,t) \in S(t)$, it shows that

(BS) implies (BS')

We have asserted that (BS') holds and therefore Theorem 9.3.1 can be applied to establish the pseudo-Lipschitz property of S^- and F^-in the following lemma.

Lemma 9.5.9 Assume that f and g satisfy (A1) and that (A2), (A3) and (NE) hold. Then for almost all $t \in [a,b]$ and all $x_1, x_2 \in int\, X(t)$ we have

$$S^-(t,x_1) \subset S^-(t,x_2) + M(t)k_x^g(t)\,|x_2 - x_1|\,\overline{B}$$

and

$$F^-(t,x_1) \subset F^-(t,x_2) + [k_x^f(t) + M(t)k_u^f(t)k_x^g(t)]\,|x_2 - x_1|\,\overline{B}.$$

The following proposition now sums up the previously obtained results into one statement on the existence of minimizers for a control problem not necessarily formulated as a differential inclusion problem. See for comparison Propositions 2.6.1 and 2.6.2 in [75].

Proposition 9.5.10 Assume that f and g satisfy (A1) and that (A2), (A3), (NE), (C) hold, the set $E \subset \mathbb{R}^n \times \mathbb{R}^n$ is closed and the function $l : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz continuous. If the control system (C) has a feasible solution (x,u), then the following optimal control problem

$$\begin{cases} \text{Minimize } l(x(a),x(b)) & \\ \text{subject to} & \\ \dot{x}(t) = f(t,x(t),u(t)), & \text{a.e. } t \in [a,b] \\ g(t,x(t),u(t)) \le 0, & \text{a.e. } t \in [a,b] \\ u(t) \in U, & \text{a.e. } t \in [a,b] \\ (x(a),x(b)) \in E, & \end{cases}$$

has a minimizer.

9.6 Conclusion

We started with a "conventional" mixed-constrained control system (C) (formulated in terms of functional relations) and introduced the multifunction F^- incorporating the state constraint and, in consequence, the notion of F^- feasible trajectories. Via this approach results which already exist for the class of differential inclusion problems become applicable to (C).

Following the approach outlined in [75] we investigated the compactness properties of F^- feasible trajectories and derive based on this an existence result for a solution. An important precondition, we made a study of the convexity properties of $F^-(t,x)$ in relation to the convexity of multifunctions $F(t,x)$ and $S^-(t,x)$.

The findings of this chapter were published as [26] in *Set-Valued and Variational Analysis*.

Final Conclusions and Future Work

Research in optimal control is often motivated by a given real-life model or a particular class of problems. Advances may lie in, for example, formulation of necessary conditions of optimality for a problem in question; in asserting regularity of the adjoint multiplier or in proving that the set of possible candidates for an optimal solution, described by the necessary coditions, is nonempty and the problem does have, in fact, an optimal solution.

The problems presented in the second (main) part of this thesis follow this pattern. The subject of Chapters 5 and 7 is the compartmental SEIR model of [61], of epidemiological background with smooth data, while Chapters 8 and 9 formulate the necessary coditions for differential-algebraic control problems, and control problems in terms of differential inclusions, respectively, both appealing to the nonsmooth maximum principle with mixed state constraints as in [20].

We derived the necessary coditions for the state constrained SEIR problem with L^2 cost and validated them against the computed ones. Hoewever, we could assert the regularity of the minimizer only on the interval $[0, T)$. We tried to remedy this shortcoming by introducing an exact penalization problem which, under an additional hypothesis, would deliver us the desired measure-free, absolutely continuous adjoint arc. However, we could not verify this hypothesis for our given SEIR problem. It remains an open question if there exists another, easier-to-verify criterion regarding neighbouring feasible trajectories of the state constrained and the exact penalization problem.

The SEIR problem with mixed contraint and the alternative L^1 cost became a very interesting topic that again allowed us to verify the optimal solution computationally. It is a subject of future work to also compute the second-order sufficient conditions of optimality.

Under some smoothness assumptions we introduce necessary conditions for differential-algebraic problems which could be advantageous in praxis since no implicit function theorem, commonly used otherwise, was required. Future research on DAE problems includes second-order sufficient conditions or also goes into direction of higher index problems.

Bibliography

[1] A. V. Arutyunov, "Optimality Conditions. Abnormal and Degenerate Problems," 1^{st} edition, Kluwer Academic Publishers, Dordrecht, 2000.

[2] A. V. Arutyunov and S. M. Aseev, *State Constraints in Optimal Control. The Degeneracy Phenomenon*, Systems & Control Letters, **26** (1995), 267–273.

[3] A. V. Arutyunov and D. Yu. Karamzin, *Necessary conditions for a weak minimum in an optimal control problem with mixed constraints*, Differ. Equ. **41** (2005), 1532–1543.

[4] J. P. Aubin and H. Frankowska, " Set-Valued Analysis," Birkhäuser, Boston, 1990.

[5] D. P. Bertesekas, " Dynamic Programming and Optimal Control," Athena Scientific, 1995.

[6] P. Bettiol, H. Frankowska, 2007. *Normality of the maximum principle for nonconvex constrained Bolza problems*, Journal of Differential Equations, 243, pp. 256-269.

[7] P. Bettiol, H. Frankowska and R.B. Vinter, 2012. L^∞ *estimates on trajectories confined to a closed subset*, J. Differential Equations, 252, pp. 1912-1933.

[8] P. Bettiol, A. Bressan and R.B. Vinter, 2010. *On trajectories satisfying a state constraint:* $W^{1,1}$ *estimates and counter-examples*, SIAM J. Control and Optimization, 48 , pp. 4664–4679.

[9] M.H.A. Biswas, L.T. Paiva and MdR de Pinho, "A SEIR model for control of infectious diseases with constraints," Mathematical Biosciences and Engineering, **11 (4)** (2014), 761-784.

[10] C. Castaing and M. Valadier, "Convex Analysis and Measurable Multifunctions," Springer Verlag, New York, 1977.

[11] L. Cesari, " Optimization-Theory and applications, Problems with ordinary differential equations," Applications of Mathematics, **17** , Springer-Verlag, New York, 1983.

[12] D. L. Cohn, "Measure Theory", Birkhäuser, 2013.

[13] F. Clarke, " Optimization and Nonsmooth Analysis," John Wiley, New York, 1983.

[14] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern and P. R. Wolenski, "Nonsmooth Analysis and Control Theory," Springer-Verlag, New York, 1998.

[15] F. Clarke, "Necessary conditions in dynamic optimization", Mem. Amer. Math. Soc., 2005.

[16] F. Clarke, *The maximum principle in optimal control, then and now*, Control Cybernet., **24** (2005) 709–722.

[17] F. Clarke, *Nonsmooth Analysis in Systems and Control Theory*, Encyclopedia of Complexity and System Science, Springer (2009), 6271–6285.

[18] F. Clarke, "Functional Analysis, Calculus of Variations and Control," Springer-Verlag, London, 2013.

[19] F. Clarke and MdR de Pinho, *The nonsmooth maximum principle*, Control Cybernet., **38** (2009) 1151–1167.

[20] F. Clarke and MdR de Pinho, *Optimal control problems with mixed constraints*, SIAM J. Control Optim., **48** (2010) 4500–4524.

[21] MdR de Pinho, *Mixed constrained control problems*, Journal of Mathematical Analysis and Applications., **278** (2003), 293–307.

[22] MdR de Pinho and R. B. Vinter, *An Euler-Lagrange inclusion for optimal control problems*, IEEE Trans. Automat. Control, **40** (1995), 1191-1198.

[23] MdR de Pinho and R. Vinter, Necessary conditions for optimal control problems involving nonlinear differential algebraic equations, J. Math. Anal. Appl., 21 (1997), pp. 493–516.

[24] MdR de Pinho, M. M. A. Ferreira, and F. A. C. C. Fontes, *An Euler-Lagrange inclusion for optimal control problems with state constraints*, . Dynam. Control Systems **8** (2002), 23–45.

[25] MdR de Pinho, M. M. A. Ferreira, and F. A. C. C. Fontes, *Unmaximized inclusion necessary conditions for nonconvex constrained optimal control problems*, ESAIM Control Optim. Calc. Var., **11** (2005) 614–632.

[26] MdR de Pinho and I. Kornienko, *Differential Inclusion Approach for Mixed Constrained Problems Revisited*, Set-Valued and Variational Analysis, published online 17 Jan 2015.

[27] MdR de Pinho, I. Kornienko and H. Maurer *Optimal Control of a SEIR Model with Mixed Constraints adn L^1 Cost*, CONTROLO2014 - Proceedings of the 11th Portuguese Conference on Automatic Control 2014.

[28] MdR de Pinho and M.M. Ferreira, *Notes on Nonsmooth Analysis and Optimal Control*, Summer School on Optimal Control and Optimization Theory, Martin-Luther-Universität Halle-Wittenberg, 2002

[29] MdR de Pinho, M.M. Ferreira, U. Ledzewicz and H. Schaettler, *A model for cancer chemotherapy with state-space constraints*, Nonlinear Analysis, **63** (2005), e2591–e2602.

[30] MdR de Pinho, P. Loewen and G. N. Silva, *A weak maximum principle for optimal control problems with nonsmooth mixed constraints*, Set-Valued and Variational Analysis, **17** 2009, 203–221.

[31] E.N. Devdaryani and Y. S. Ledyaev , *Maximum principle for implicit control systems*, Appl. Math. Optim., **40** (1999) 79–103.

[32] A. V. Dmitruk, *Maximum principle for the general optimal control problem with phase and regular mixed constraints*, Comput. Math. Model., **4** (1993) 364–377.

[33] A. Ya. Dubovitskii and A.A. Milyutin, "Extremum problems under constraints", Dokl. Akad. Nauk SSSR, **149** (1963), 1065–1089.

[34] P. Falugi, E. Kerrigan and E. van Wyk, *Imperial College London Optimal Control Software User Guide (ICLOCS)*, Department of Electrical and Electronic Engineering, Imperial College London, London, England, UK, 2010.

[35] M. M. A.Ferreira, F. A. C. C. Fontes and R. B. Vinter, *Nondegenerate Necessary Conditions for Nonconvex Optimal Control Problems with State Constraints*, Journal of Mathematical Analysis and Applications, **233**,(1999), 116–129.

[36] G. B. Folland, "Real Analysis: Modern Techniques and Their Applications," 2nd Edition, A Wiley-Interscience Publication, John Wiley and Sons, New York, 1999.

[37] R. Fourer, D.M. Gay and B.W. Kernighan, "AMPL: A Modeling Language for Mathematical Programming," Duxbury Press, Brooks–Cole Publishing Company, 1993.

[38] H. Frankowska, *Regularity of minimizers and of adjoint states for optimal control problems under state constraints*. J. Convex Analysis, 13, (2006), pp. 299 – 328.

[39] M. Gerdts, *Local minimum principle for optimal control problems subject to differential-algebraic equations of index two*, J. Optim. Theory Appl., 130 (2006), pp. 441–460.

[40] M. Gerdts, "Optimal Control of ODEs and DAEs," De Gruyters, Berlin, 2012.

[41] E. Giner, *Local Minimizers of Integral Functionals are Global Minimizers*, Proceedings of the American Mathematical Society, **123** (3), (1995), 755–757.

[42] R. F. Hartl, S. P. Sethi and R. G. Vickson *A Survey of the Maximum Principles for Optimal Control Problems with State Constraints*, SIAM Review, **37**, (2), (1995) 181–281.

[43] M. R. Hestenes, " Calculus of Variations and Optimal Control Theory," John Wiley, New York, 1966.

[44] A. D. Ioffe and V. M. Tikhomirov, " Theory of extremal problems," North-Holland Pub. Co., 1979.

[45] I. Kornienko and MdR de Pinho, 2013. *Properties of control systems with mixed constraints in the form of inequalities*. Internal Report, ISR, DEEC, FEUP, 2013.

[46] I. Kornienko, M. Gerdts, and MdR de Pinho, *New version of necessary conditions for optimal control problems with differential algebraic equations*, Proceedings of MTNS 2012, Melbourne, Australia, 2012.

[47] I. Kornienko and MdR de Pinho, *On the first order state constrained optimal control problems*, Proceedings of MTNS 2014, Groningen, Netherlands, 2014.

[48] I. Kornienko, L. T. Paiva, and MdR de Pinho, *Introducing State Constraints in Optimal Control for Health Problems*, Proceedings of CETC 2013 Lisbon, Procedia Technology, **17**, (2014), 415–422

[49] H. W. Kuhn and A. W. Tucker, *Nonlinear Programming*, Proc. Second Berkeley Symp. on Math. Statist. and Prob.,(J. Neyman, Ed.). Univ. of Calif. Press, (1951), 481–492.

[50] P. Kunkel and V. Mehrmann, *Optimal control for unstructured nonlinear differential-algebraic equations of arbitrary index*, Math. Control Signals Syst., 22 (2008) pp. 227-269.

[51] P. D. Loewen and R. T. Rockafellar, *New Necessary Conditions for the Generalized Problem of Bolza*, SIAM J. Control and Optimization-Preprint, **1** (1995) .

[52] S. O. Lopes, "Nondegenerate forms of the Maximum Principle for Optimal Control Problems with State Constraints," PhD Thesis, University of Minho, Portugal. 2008.

[53] S. O. Lopes and F.A.C.C. Fontes, *On Stronger Forms of First-Order Necessary Conditions of Optimality for State-Constrained Control Problems* International Journal of Pure and Applied Mathematics, **49**(4) (2008), 459–466.

[54] D. G. Luenberger, " Optimization by Vector Space Methods," John Wiley and Sons, Inc., New York, 1969.

[55] H. Maurer and H.J. Oberle, *Second order sufficient conditions for optimal control problems with free final approach: The Riccati approach*, SIAM J. Control Optm., **41** (2002), 380–403.

[56] H. Maurer and N.P. Osmolovskii, *Second-order conditions for optimal control problems with mixed control-state constraints and control appearing linearly*, Proceedings of the 52nd IEEE Conference on Control and Design (CDC 2013), Firenze, (2013), 514–519.

[57] H. Maurer and S. Pickenhain, *Second order sufficient conditions for optimal control problems with mixed control-state constraints*, J. Optim. Theory Appl., **86** (1995), 649–667.

[58] A. A. Milyutin and N. P. Osmolovskii, "Calculus of Variations and Optimal Control," Translations of Mathematical Monographs 180, American Mathematical Society, Providence, Rhode Island, 1998.

[59] B. Mordukhovich, "Variational analysis and generalized differentiation. Basic Theory", Springer-Verlag, Berlin, 2006.

[60] I. P. Natanson, "Theory of Functions of A Real Variable", Frederick Ungar Publishing, 1964.

[61] R.M. Neilan and S. Lenhart, *An Introduction to Optimal Control with an Application in Disease Modeling*, DIMACS Series in Discrete Mathematics, **75** (2010), 67–81.

[62] L. W. Neustadt, "Optimization, A Theory of Necessary Conditions," Princeton University Press, New Jersey, 1976.

[63] N. P. Osmolovskii and H. Maurer, *Applications to Regular and Bang-Bang Control: Second-Order Necessary and Sufficient Optimality Conditions in Calculus of Variations and Optimal Control*, SIAM Advances in Design and Control, **24**, 2013.

[64] L.T. Paiva, *Optimal Control in Constrained and Hybrid Nonlinear Systems*, Project Report (2013), http://paginas.fe.up.pt/~faf/ProjectFCT 2009/report.pdf.

[65] Z. Páles and V. Zeidan, *Optimal control problems with set-valued control and state constraints*, SIAM J. Optim., **14** (2003) 334-358.

[66] H. J. Pesch and M. Plail "The Maximum Principle of optimal control: A history of ingenious ideas and missed opportunities", Control and Cybernetics, **38** (2009), 973–995.

[67] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze and E. F. Mischenko, "The Mathematical Theory of Optimal Multiprocesses," John Wiley, New York, 1962.

[68] F. Rampazzo, R. Vinter, *A theorem on existence of neighbouring trajectories satisfying a state constraint, with applications to optimal control*, IMA J. Math. Control & Information, **16 (4)**, (1999) 335–351.

[69] F. Rampazzo and R. Vinter, *Degenerate Optimal Control Problems with State Constraints*, SIAM J. Control Optim., **39**(4), (2000) 989-1007.

[70] R.T. Rockafellar and B. Wets, "Variational Analysis," Springer Verlag, Berlin, 1998.

[71] T. Roubicek and M. Valasek, *Optimal control of causal differential-algebraic systems*, (2002) J. Math. Anal. Appl., 269, pp. 616-641.

[72] H. Schaettler and U. Ledzewicz, "Geometric Optimal Control. Theory, Methods and Examples", Springer, New York, 2012.

[73] I. Shvartsman, R. Vinter, *Regularity properties of optimal controls for problems with time-varying state and control constraints*, Nonlin. Anal.: Theory, Methods & Applications, **65 (2)**, (2006), 448–474.

[74] R.B. Vinter and G. Pappas, *A maximum principle for nonsmooth optimal-control problems with state constraints*, J. Math. Anal. Appl., **89** (1982) 212–232.

[75] R.B. Vinter, "Optimal Control," Birkhäuser, Boston, 2000.

[76] A. Wächter and L. T. Biegler, *On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming*, Mathematical Programming, **106** (2006), 25–57.

[77] P. D. Woodford, "Optimal Control for Nonsmooth Dynamic Systems," Ph D Thesis, Centre for Process Systems Engineering and Department of Electrical and Electronic Engineering, Imperial College of Science, Technology and Medicine, London SW7 2BY, 1997.